Contents

Original Papers

Challenges for Medical Students in Applying Ethical Principles to Allocate Life-Saving Medical Devices During the COVID-19 Pandemic: Content Analysis (e52711)
Hsing-yen Hsieh, Chyi-her Lin, Ruyi Huang, Guan-chun Lin, Jhen-Yu Lin, Clydie Aldana. ................................................................. 5

A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study (e53961)
Friederike Holderried, Christian Stegemann–Philipp, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling. ................................................................. 14

Unpacking the Experiences of Health Care Professionals About the Web-Based Building Resilience At Work Program During the COVID-19 Pandemic: Framework Analysis (e49551)
Wei Ang, Zhi Lim, Siew Lau, Jie Dong, Ying Lau. .................................................................................................................. 61

Occupational Therapy Students’ Evidence-Based Practice Skills as Reported in a Mobile App: Cross-Sectional Study (e48507)
Susanne Johnson, Birgitte Espehaug, Lillebeth Larun, Donna Giliska, Nina Olsen. ........................................................................ 74

Evaluating the Impact of the National Health Service Digital Academy on Participants’ Perceptions of Their Identity as Leaders of Digital Health Change: Mixed Methods Study (e46740)
Amish Acharya, Ruth Black, Alisdair Smithies, Ara Darzi. .......................................................... 83

Development of a Clinical Simulation Video to Evaluate Multiple Domains of Clinical Competence: Cross-Sectional Study (e54401)
Kiyoshi Shikino, Yuji Nishizaki, Sho Fukui, Daiki Yokokawa, Yu Yamamoto, Hiroyuki Kobayashi, Taro Shimizu, Yasuharu Tokuda. .................................................................................. 96

Design and Development of Learning Management System Huemul for Teaching Fast Healthcare Interoperability Resource: Algorithm Development and Validation Study (e45413)
Sergio Guinez-Molinos, Sonia Espinoza, Jose Andrade, Alejandro Medina. .......................................................... 107

The Effects of Immersive Virtual Reality–Assisted Experiential Learning on Enhancing Empathy in Undergraduate Health Care Students Toward Older Adults With Cognitive Impairment: Multiple-Methods Study (e48566)
Justina Liu, Pui Mak, Kitty Chan, Daphne Cheung, Kin Cheung, Kenneth Fong, Patrick Kor, Timothy Lai, Tulio Maximo. .......................................................... 120

Measuring e-Professional Behavior of Doctors of Medicine and Dental Medicine on Social Networking Sites: Indexes Construction With Formative Indicators (e50156)
Marko Mareli, Ksenija Klasni, Tea Vukuši, Rukavina. .................................................................................................................. 136
Pure Wisdom or Potemkin Villages? A Comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 Style Questions: Quantitative Analysis (e51148)  
Leonard Knoedler, Michael Alftertshofer, Samuel Knoedler, Cosima Hoch, Paul Funk, Sebastian Cotofana, Bhagvat Maheta, Konstantin Frank, Vanessa Brébant, Lukas Prantl, Philipp Lamby. ................................................................. 164

Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects (e51247)  
Lukas Weidener, Michael Fischer. ........................................................................... 174

Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study (e51308)  
Amanda Zaleski, Rachel Berkowsky, Kelly Craig, Linda Pescatello. ............................ 192

The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators’ Perspectives (e47339)  
Yaser Al-Worafi, Khang Goh, Andi Hermansyah, Ching Tan, Long Ming. .................. 207

A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology–Head and Neck Surgery Certification Examinations: Performance Study (e49970)  
Cai Long, Kayle Lowe, Jessica Zhang, André Santos, Alaa Alanazi, Daniel O’Brien, Erin Wright, David Cote. ................................................................. 214

Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study (e50842)  
Firas Haddad, Joanna Saade. .................................................................................... 237

Evaluation of ChatGPT’s Real-Life Implementation in Undergraduate Dental Education: Mixed Methods Study (e51344)  
Argyro Kavadella, Marco Dias da Silva, Eleftherios Kaklamanos, Vasileios Stamatopoulos, Kostis Giannakopoulos. ................................................................. 244

Increasing Realism and Variety of Virtual Patient Dialogues for Prenatal Counseling Education Through a Novel Application of ChatGPT: Exploratory Observational Study (e50705)  
Megan Gray, Austin Baird, Taylor Sawyer, Jasmine James, Thea DeBroux, Michelle Bartlett, Jeanne Krick, Rachel Umoren. ............................................................... 258

Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study (e50965)  
Annika Meyer, Janik Riese, Thomas Streichert. .......................................................... 268

Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study (e48514)  
Peng Yu, Changchang Fang, Xiaolin Liu, Wanying Fu, Jitao Ling, Zhiwei Yan, Yuan Jiang, Zhengyu Cao, Maxiong Wu, Zhiteng Chen, Wengen Zhu, Yuling Zhang, Ayiguli Abudukeremu, Yue Wang, Xiao Liu, Jingfeng Wang. ......................................................... 280

Cocreating an Automated mHealth Apps Systematic Review Process With Generative AI: Design Science Research Approach (e48949)  
Guido Giunti, Colin Doherty. .................................................................................... 289

Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models (e51391)  
Tassallah Abdullahi, Ritambhara Singh, Carsten Eckhoff. ........................................... 299

Exploring the Feasibility of Using ChatGPT to Create Just-in-Time Adaptive Physical Activity mHealth Intervention Content: Case Study (e51426)  
Amanda Willms, Sam Liu. ....................................................................................... 315
AI Education for Fourth-Year Medical Students: Two-Year Experience of a Web-Based, Self-Guided Curriculum and Mixed Methods Study (e46500)
Areeba Abid, Avinash Murugan, Imon Banerjee, Saptarshi Purkayastha, Hari Trivedi, Judy Gichoya. 327

Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard (e51523)
Faiza Farhat, Beenish Chaudhry, Mohammad Nadeem, Shahab Sohail, Dag Madsen. 334

Viewpoints

Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education (e55368)
Lukas Weidener, Michael Fischer. 32

Patients, Doctors, and Chatbots (e50869)
Thomas Erren. 151

Generative Language Models and Open Notes: Exploring the Promise and Limitations (e51183)
Charlotte Blease, John Torous, Brian McMillan, Maria Häggglund, Kenneth Mandl. 155

Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project (e51388)
Nicholas Kuo, Oscar Perez-Concha, Mark Hanly, Emmanuel Mnatzaganian, Brandon Hao, Marcus Di Sipio, Guolin Yu, Jash Vanjara, Ivy Valerie, Juliana de Oliveira Costa, Timothy Churches, Sanja Lujic, Jo Hegarty, Louisa Jorm, Sebastiano Barbieri. 222

Using ChatGPT-Like Solutions to Bridge the Communication Gap Between Patients With Rheumatoid Arthritis and Health Care Professionals (e48989)
Chih-Wei Chen, Paul Walter, James Wei. 310

Tutorial

Sharing Digital Health Educational Resources in a One-Stop Shop Portal: Tutorial on the Catalog and Index of Digital Health Teaching Resources (CIDHR) Semantic Search Engine (e48393)
Julien Grosjean, Arriel Benis, Jean-Charles Dufour, Émilene Lejeune, Flavien Disson, Badisse Dahanma, Hélène Cieslik, Romain Léguillon, Matthieu Faure, Frank Dufour, Pascal Staccini, Stéfan Darmoni. 47

Corrigenda and Addenda

Correction: How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment (e57594)
Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Taylor, David Chartash. 346

Editorial

ChatGPT in Medical Education: A Precursor for Automation Bias? (e50174)
Tina Nguyen. 348
Short Paper

Impact of the COVID-19 Pandemic on Medical Grand Rounds Attendance: Comparison of In-Person and Remote Conferences (e43705)
Ken Monahan, Edward Gould, Todd Rice, Patty Wright, Eduard Vasilevskis, Frank Harrell, Monique Drago, Sarah Mitchell. 354

Research Letter

Using AI Text-to-Image Generation to Create Novel Illustrations for Medical Education: Current Limitations as Illustrated by Hypothyroidism and Horner Syndrome (e52155)
Ajay Kumar, Pierce Burr, Tim Young. 362
Challenges for Medical Students in Applying Ethical Principles to Allocate Life-Saving Medical Devices During the COVID-19 Pandemic: Content Analysis

Hsing-yen Hsieh¹, PhD; Chyi-her Lin¹,², MD; Ruyi Huang¹,³,⁴,⁵, MPH, MD; Guan-chun Lin¹, PhD; Jhen-Yu Lin³, MS; Clydie Aldana¹, BS

¹School of Medicine for International Students, College of Medicine, I-Shou University, Kaohsiung, Taiwan
²Department of Pediatrics, E-Da Hospital, Kaohsiung, Taiwan
³Holistic Medicine, Department of Family and Community Medicine, E-Da Hospital, Kaohsiung, Taiwan
⁴Data Science Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan
⁵Division of Family Medicine, Taipei Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, and School of Medicine, Tzu Chi University, Hualien, Taiwan

Corresponding Author:
Ruyi Huang, MPH, MD
School of Medicine for International Students
College of Medicine
I-Shou University
8 Yida Rd
Yanchao District
Kaohsiung, 82445
Taiwan
Phone: 886 7 615 0011 ext 2547
Fax: 886 7 615 0940
Email: ruyi@mail.harvard.edu

Abstract

Background: The emergence of the COVID-19 pandemic has posed a significant ethical dilemma in the allocation of scarce, life-saving medical equipment to critically ill patients. It remains uncertain whether medical students are equipped to navigate this complex ethical process.

Objective: This study aimed to assess the ability and confidence of medical students to apply principles of medical ethics in allocating critical medical devices through the scenario of virtual patients.

Methods: The study recruited third- and fourth-year medical students during clinical rotation. We facilitated interactions between medical students and virtual patients experiencing respiratory failure due to COVID-19 infection. We assessed the students’ ability to ethically allocate life-saving resources. Subsequently, we analyzed their written reports using thematic analysis to identify the ethical principles guiding their decision-making.

Results: We enrolled a cohort of 67 out of 71 medical students with a mean age of 34 (SD 4.7) years, 60% (n=40) of whom were female students. The principle of justice was cited by 73% (n=49) of students while analyzing this scenario. A majority of them expressed hesitancy in determining which patient should receive life-saving resources, with 46% (n=31) citing the principle of nonmaleficence, 31% (n=21) advocating for a first-come-first-served approach, and 25% (n=17) emphasizing respect for patient autonomy as key influencers in their decisions. Notably, medical students exhibited a lack of confidence in making ethical decisions concerning the distribution of medical resources. A minority, comprising 12% (n=8), proposed the exploration of legal alternatives, while 4% (n=3) suggested medical guidelines and collective decision-making as potential substitutes for individual ethical choices to alleviate the stress associated with personal decision-making.

Conclusions: This study highlights the importance of improving ethical reasoning under time constraints using virtual platforms. More than 70% of medical students identified justice as the predominant principle in allocating limited medical resources to critically ill patients. However, they exhibited a lack of confidence in making ethical determinations and leaned toward principles such as nonmaleficence, patient autonomy, adherence to legal and medical standards, and collective decision-making to mitigate the pressure associated with such decisions.
virtual patient; virtual patients; medical resources distribution; medical ethical education; COVID-19 pandemic; ethics; medical student; medical students; medical ethics; decision-making; ethical dilemma; simulation; reasoning; decision support; medical guideline; medical guidelines; medical devices; medical device; life-saving; thematic analysis; virtual platform

**Introduction**

The COVID-19 pandemic has caused millions of deaths and countless hospitalizations worldwide owing to critical conditions caused by the virus [1]. This has raised the ethical dilemma of allocating scarce life-saving devices to critically ill patients [2-5].

Physicians often make clinical decisions based on scientific evidence to avoid moral distress [3,6,7]. However, clinical decisions may have to be made under time constraints. Preparing physicians to apply appropriate ethical principles, have self-confidence in making choices, and prevent moral trauma has become essential during the pandemic [8].

The principles of autonomy, justice, beneficence, and nonmaleficence commonly serve as guiding references for allocating scarce medical resources [9]. However, these principles have multiple interpretations when facing limited resources and can be based on utilitarianism, egalitarianism, or deontology [10]. Utilitarianism believes that the primary obligation is not to treat people equally, but to maximize the greatest amount of happiness for the greatest number of people; the best actions would be based on what brings the best benefit. By contrast, egalitarianism upholds the rights and interests of individuals, which should be equally protected [10]. Deontology judges the morality of choices by its conformity with a moral norm [11], regardless of its consequences. Persad et al [12] present a comprehensive framework for the allocation of scarce medical resources grounded in the core principles of autonomy, justice, beneficence, and nonmaleficence. Their framework encompasses 4 distinct ethical value categories, including equal treatment, prioritization of the most vulnerable, maximizing overall benefits, and recognition of social usefulness. Within each category, 2 competing ethical principles emerge, yielding a total of 8 subprinciples that provide detailed guidance aligned with the overarching ethical values [12]. The core values or principles that medical students prefer or overlook when facing ethical dilemmas are unclear and require further study.

The School of Medicine for International Students at I-Shou University has a 4-year Doctor of Medicine program that collaborates with the Ministry of Foreign Affairs and enrolls college graduates from countries with official diplomatic ties to Taiwan. Due to the limited medical resources in such students’ home countries, they may face the challenge of a shortage of life-saving medical facilities in clinical practice. Therefore, equipping them with the knowledge and skills to allocate life-saving medical devices to critically ill patients, based on reasonable principles of medical ethics, is crucial. The use of virtual patients for teaching medical humanities may strengthen the effectiveness of medical ethics education [13,14]. Considering the challenges imposed by the COVID-19 pandemic, this solution aims to offer a secure and personalized training environment, transcending the boundaries of time and space. By doing so, students can become fully engrossed in virtual scenarios, enriching their learning experiences.

The objective of this study was to assess the ability and confidence of medical students to apply principles of medical ethics in allocating critical medical devices through the scenario of virtual patients.

**Methods**

**Study Design**

We designed a virtual scenario and asked medical students to allocate lifesaving medical devices to only 1 patient. In this scenario, a 62-year-old COVID-19-infected patient with respiratory failure was admitted to the intensive care unit. Medical students were instructed to interview a virtual patient and review the patient’s laboratory and imaging findings. They then were asked to make clinical diagnoses and adopt appropriate ethical principles to determine whether to remove the extracorporeal membrane oxygenation (ECMO) device from an 80-year-old patient currently using it and reallocate it to the new younger patient. After making their decision, the students were requested to write a short essay addressing the ethical conflicts they encountered in making the choice.

**Ethical Considerations**

We explained the rationale for this qualitative study and recruited third- and fourth-year medical students from the School of Medicine for International Students Program when they undertook clinical rotation at the hospital. All participants completed the virtual clinical scenarios within 4 hours in May 2021, during the COVID-19 pandemic in Taiwan, after signing an informed consent form. This study was approved by the E-Da Hospital Institutional Review Board (no. EMRP05109N and EMRP04111N), and the data were not identifiable. The teaching and evaluation of students were not affected by whether they participated in the research.

**Case Scenario**

Leona is a 62-year-old retired woman. She had been well without any underlying disease until recently being diagnosed with COVID-19 pneumonitis. Her lung condition continuously deteriorated, and ECMO was the last resort to support her tissue oxygenation. However, the only available ECMO machine was currently being used by an 80-year-old patient with multiple chronic illnesses who remained unstable after receiving ECMO treatment. In this study, we may need to remove the ECMO from Leona.

The students were given the above scenario to assess and answer relevant questions. One of the questions was “Will you continue to let the 80-year-old patient use the ECMO, or let Leona use it?”
the ECMO instead? Please explain your decision and your reasons to support it.”

The medical students could use the 4 principles of medical ethics or base their responses on their individual analytical perspectives and reasoning for the allocation of limited medical resources.

Data Analysis
Age (>25 vs ≤25 years) and sex (male vs female) served as basic demographic variables, with the age of 25 years as a threshold of maturity. Grade (third vs fourth year) represented differences in clinical exposure experiences [15]. Textual content analysis was performed by 2 of the authors to search for keywords and summarize the students’ responses independently. The keywords were encoded and categorized for both quantitative and qualitative analyses. We used the principles of summative content analysis, which combines the quantitative counting of specific content or words or terms with latent content analysis to identify and categorize their meanings. In brief, we created a new coding category for any newly introduced terms in the assignment, and then assessed conceptual similarities to determine whether to further organize these codes into additional categories with appropriate names.

The qualitative analysis consisted of the following steps:

1. The coding items included the final decision of the students (for whom to use), which core medical ethical principles were applied with various degrees in their choices, and whether viewpoints other than ethics, such as medical guidelines or legislation, were mentioned.

2. The reasons for the students’ final decisions were classified according to the patient they selected, either the 62-year-old younger patient or the 80-year-old patient with multiple comorbidities. Our analysis focused on encoding the ethical justifications provided by the medical students to support their final decisions. We omitted considerations related to their alternative choices during the decision-making process.

3. The classification of reasoning for those who made a decision was primarily based on the students’ understanding and interpretations in their essays, which Persad et al [12]

mentioned were equality, vulnerability, maximizing the quality of life, and contribution to society. The original resource allocation principles were designed for the distribution of medical supplies among a group of individuals. However, the present case pertains to the treatment decision for an individual patient, further complicated by the fact that one patient had already been put on a ventilator. By contextualizing the principles within the framework of the present case, we eliminated the applicability of 4 subprinciples: lottery, saving the most lives, reciprocity, and giving priority to the worst off (ie, sickest first).

4. If students displayed reluctance in making a choice, we also coded their explanations for the perception that ethical decision-making might not be suitable, categorizing these explanations as “undetermined” or “both unqualified.”

5. The main reasons for the students’ final decisions were classified into medical, legal, and ethical perspectives.

6. The coding process was independently judged by 2 researchers with expertise in qualitative research. Any inconsistencies in coding were resolved by reviewing the classification descriptions to refine the precision of category definitions and revisiting the context to ensure accurate coding.

Results

Student Demographics
From 2021 to 2022, a total of 71 international third- and fourth-year clinical medical students who were facing the COVID-19 pandemic most significantly were enrolled. Of these, 67 students (33 third-year and 34 fourth-year students) from 12 countries participated in the study. Because 4 fourth-year medical students did not participate, the response rate was 94%. Overall, 40 (60%) participants were female and 61 (91%) were older than 25 years. Most medical students were from the Kingdom of Eswatini, accounting for 48% (n=32) of the total group (Table 1 and Multimedia Appendix 1).
Table 1. Basic information of the students.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Medical students (n=67), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>27 (40)</td>
</tr>
<tr>
<td>Female</td>
<td>40 (60)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>&gt;25</td>
<td>61 (91)</td>
</tr>
<tr>
<td>≤25</td>
<td>6 (9)</td>
</tr>
<tr>
<td><strong>Seniority year</strong></td>
<td></td>
</tr>
<tr>
<td>Third</td>
<td>33 (49)</td>
</tr>
<tr>
<td>Fourth</td>
<td>34 (51)</td>
</tr>
<tr>
<td><strong>Country of origin</strong></td>
<td></td>
</tr>
<tr>
<td>The Kingdom of Eswatini</td>
<td>32 (48)</td>
</tr>
<tr>
<td>Saint Lucia</td>
<td>7 (10)</td>
</tr>
<tr>
<td>Belize</td>
<td>7 (10)</td>
</tr>
<tr>
<td>Kiribati</td>
<td>5 (7)</td>
</tr>
<tr>
<td>Honduras</td>
<td>3 (4)</td>
</tr>
<tr>
<td>The Marshall Islands</td>
<td>3 (4)</td>
</tr>
<tr>
<td>Saint Kitts and Nevis</td>
<td>3 (4)</td>
</tr>
<tr>
<td>Paraguay</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Saint Vincent &amp; The Grenadines</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Palau</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Haiti</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Solomon Islands</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>

Choosing the Best Candidate for ECMO Allocation

Of the 67 participating students, age group (<25 vs ≥25 years old), sex (male vs female), and seniority year (third vs fourth year) did not affect patient selection preferences, and a larger proportion of students from Eswatini (21/32, 66%) selected the 80-year-old patient for ECMO compared to the rest of the students (39/67, 58%). The majority of students decided to continue treating the 80-year-old patient with ECMO (Table 2).

Additionally, 5 (8%) students argued that the medical information provided was not sufficient to make decisions that were highly dependent on factors such as the patient’s condition, the course of the disease, and legal requirements. One student (1%) suggested that, in accordance with medical guidelines, neither patient met the conditions to be a candidate for ECMO. A possible reason for them to abstain from decision-making could be the pressure they experienced while facing an ethical dilemma. As one student (no. 16) stated:

*Doctors should not take the treatment away of one person and give it to another, regardless of the odds of survival rate of these two patients, because it means that we are taking the role of God, deciding who lives and who dies.*

Another student (no. 20) stated:

*I don’t believe I have the right to decide who is more deserving or who needs this equipment more.*

Table 2. Choosing the most suitable patient for extracorporeal membrane oxygenation treatment.

<table>
<thead>
<tr>
<th>Patient selected</th>
<th>Students (n=67), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80-year-old</td>
<td>39 (58)</td>
</tr>
<tr>
<td>62-year-old</td>
<td>22 (33)</td>
</tr>
<tr>
<td>Undetermined</td>
<td>5 (8)</td>
</tr>
<tr>
<td>Both unqualified</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>
Students’ Perspective of Allocating Limited Resources

Building upon the framework proposed by Persad et al [12], this study identified 4 coding categories after excluding subprinciples that were deemed inapplicable to the current case. In accordance with the students’ final decisions regarding the most suitable recipient for ECMO, we categorized the reasons endorsed by the students (Table 3). The primary justifications for selecting an 80-year-old patient included nonmaleficence (n=31, 46%), first-come-first-served (n=21, 31%), and patient autonomy (n=17, 25%). Students grounded their decisions in 3 of the 4 ethical principles, arguing that in this particular scenario, those advocating for the principle of nonmaleficence contended that physicians lacked the authority to withdraw a life-saving device in active use. “First-come-first-served” represents 1 of the 4 interpretive angles of the justice principle from Persad’s framework. Students believed that the life of each patient held equal value, and those who received treatment first should be allowed to continue treatment. Students who mentioned patient autonomy were particularly concerned about the absence of informed consent and its potential legal implications for health care providers.

The reasons for selecting the 62-year-old patient primarily revolved around the principle of justice. The utilitarian principle of maximum benefit was the most popular: 31% (n=21) of students mentioned that medical resources should be reserved for patients who can survive the longest and have the best quality of life. When comparing who had better survival probabilities, some students suggested that medical guidelines should serve as the basis for the final decision. Overall, 10% (n=7) of students made decisions depending on who had contributed more to society as a whole, and 4% (n=3) prioritized the disadvantaged, where the disadvantaged can be interpreted as the younger patient.

Students who expressed an “undetermined” stance believed that decision-making authority should be entrusted to guidelines, which could be either principles collectively established by physicians within the hospital (n=4, 6%), hospital policies (n=4, 6%), local laws (n=4, 6%), or decisions made by the hospital’s ethics committee (n=3, 4%). Alternatively, some advocated for decisions to be made collectively by physicians within the hospital (n=1, 1%), by the patients’ families (n=1, 1%), or based on other information relevant to the patient’s condition (n=1, 1%). One student expressed a “both unqualified” position and approached the issue from a medical rather than an ethical perspective. The student asserted that, based on the guidelines, neither of the 2 patients met the criteria for usage.

Table 3. Multiple-choice analysis of the reasoning for case selection among students.

<table>
<thead>
<tr>
<th>Reasoning for selected patient</th>
<th>Students (n=67), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80-year-old</td>
<td></td>
</tr>
<tr>
<td>Nonmaleficence (physician has no right to withdraw)</td>
<td>31 (46)</td>
</tr>
<tr>
<td>Treat patients equally (first come, first served)</td>
<td>21 (31)</td>
</tr>
<tr>
<td>Patient’s autonomy (law issue)</td>
<td>17 (25)</td>
</tr>
<tr>
<td>Withdraw can’t prove 62-year-old patient’s survival</td>
<td>2 (3)</td>
</tr>
<tr>
<td>62-year-old</td>
<td></td>
</tr>
<tr>
<td>Higher survival rate, save the maximum quality of life (medical issue)</td>
<td>21 (31)</td>
</tr>
<tr>
<td>Rewarding social usefulness</td>
<td>7 (10)</td>
</tr>
<tr>
<td>Giving priority to the worst off; youngest first</td>
<td>3 (4)</td>
</tr>
<tr>
<td>Undetermined</td>
<td></td>
</tr>
<tr>
<td>Decided by medical guidelines, collective decision</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Decided by hospital</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Depend on law</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Decided by the ethics committee</td>
<td>3 (4)</td>
</tr>
<tr>
<td>Decided by 80-years-old patient’s family member</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Depend on other medical information</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Both unqualified</td>
<td></td>
</tr>
<tr>
<td>Both are unqualified for ECMO&lt;sup&gt;a&lt;/sup&gt; per guidelines</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>

<sup>a</sup>ECMO: extracorporeal membrane oxygenation.

Adequacy of Using Medical Ethical Principles

In total, 73% (n=49) of students cited the principle of justice while analyzing this case. When ethical principles were in conflict, the principle of justice was most commonly cited. The frequencies of ethical principles considered by medical students in making final decisions (coding as simple choice) were as follows: 48% (n=32) used the principle of justice, 25% (n=18) used the principle of nonmaleficence, 12% (n=8) used the principle of patient autonomy, and 9% (n=6) were unable to provide a definitive response.
Confidence in Ethical Decision-Making

Overall, 75% (n=50) of the participants analyzed the case from other perspectives, such as medicine and law, and 25% (n=18) made their final decision based on the principles mentioned in the clinical guidelines. These students were more inclined toward the scientific mode of thinking, believing that evidence-based medicine is objective and may provide clear standards that can give them a sense of security. Students no. 23 and 31, respectively, indicated the following:

*I can respond to this situation based on scientific evidence.*

*A comprehensive assessment of the pathology of the patient’s current condition and the state of illness is a major consideration in decision-making.*

For 12% (n=8) of the medical students, their final decisions were made from a legal perspective; that is, they stated that the decision should be made in accordance with the law of the state. They emphasized that physicians should protect themselves from being sued and provide decision-making authority to the patient or family. The patients or their family members should sign the emergency consent form, allowing the patient or family to participate in decision-making. As stated by student no. 40:

*Medical care providers must consider medical laws, including those for removing the machine from the patient and withholding services from patients.*

Additionally, 6% (n=4) of the medical students believed that medical institutions should provide clear guidelines or set up ethics committees to make collective decisions, thus preventing individual doctors from facing the pressure of decision-making. Student no. 18 stated:

*I will follow the organization’s code of ethics. The handling rules approved by a specific organization that will guide you in such situations so that you do not face a violation of the law.*

Discussion

Principal Findings

ECMO is recommended for severe COVID-19-related acute respiratory distress syndrome to reduce mortality [16]. Currently, there is no evidence-based ethical guidance for prioritizing ECMO when resources are limited during the COVID-19 pandemic [17]. Justice is the preferred principle in virtual settings, although students have diverse interpretations. Nearly half of the students used additional principles, such as nonmaleficence and respect for patient autonomy, to prevent further harm while making ethical decisions. Multiple perspectives were adopted by three-fourths of the students.

The context of clinical situations is important for making clinical decisions based on ethical dilemmas [18]. The use of virtual patients for medical education may strengthen the effectiveness of medical ethics education [13,14]. Using virtual patients for clinical decision-making training among international medical students offers several advantages [19-21]. It provides a safe training environment amidst the COVID-19 pandemic and allows for diverse case presentations from multiple countries and cultures [22]. The application of virtual care has flourished internationally during the post-COVID era. The Cleveland Medical Center in the United States has also explored the integration of remote and virtual health care. Medical institutions in the southern United States have proved that virtual diagnosis and treatment can alleviate caregiver burden and promote patient care [23]. Our study has provided evidence that combining virtual training with ethical reasoning in solving ethical dilemmas may present a safe environment for learning clinical decision-making and offer opportunities for improvement.

Students were asked to think about and answer questions according to the situation of the virtual patient. More than half of the students chose the oldest or the sickest patient to be the best candidate. The clinical scenario that was tested involved ex-post triage, which entails discontinuing ongoing treatment in favor of a newly arrived patient. Particularly in the context of a pandemic with limited resources (eg, ventilators), the primary objective is to maximize overall benefits for all individuals. While challenging, medical physicians may need to make the difficult decision of reallocating life-saving facilities from the most critically ill patients to those who have a higher probability of survival [5]. During a pandemic, rationing may require the withdrawal of care in order to provide ventilators to patients who are given higher priority, a reason foreign to many front-line clinicians [24]. Sharing and leveraging the diverse responses of medical students themselves can serve as a valuable reference for fostering innovative approaches in medical ethics education and facilitating ethical deliberation on challenging medical issues.

Medical students must define problems, identify potential solutions, and also inform patients about the current treatment options. The students’ understanding of patient autonomy and informed consent was superficial and formalistic; they were more concerned about obtaining consent or documents to avoid legal proceedings. Recent discussions on the principles of patient autonomy have concluded that superficial autonomy cannot guarantee patient autonomy [25-27]. Moreover, physicians should make more efforts to meet the best interests of patients [28,29]. Considering students’ diverse backgrounds, it is important to take into account their various learning styles to enhance and personalize educational materials [30].

The inability to establish a definitive ethical guideline capable of resolving issues stemming from the scarcity of medical resources underscores the complexity of the situation. Furthermore, factors such as patients possessing varying medical needs, financial capabilities to cover medical expenses, and the policies of health care institutions can all impact the ethical judgments of students [31,32]. Therefore, teachers can take the opportunity to emphasize to students that the premise of patient autonomy and informed consent is to uphold the patient’s right to live, and promoting the well-being of the patient is the core value of the principle of patient autonomy. To ensure the patient’s autonomy is respected, physicians should make decisions that benefit the patient’s overall health and care.

Students were unfamiliar with philosophical and ethical reasoning and were under pressure to make ethical decisions about allocating life-saving medical modalities. They tended to...
analyze ethical issues from both medical and legal perspectives [33,34]. Most medical students relied on objective medical guidelines, legal documents, or hospital management systems to help them make decisions while lacking life-saving medical modalities. Experts might erroneously assume that by dutifully adhering to the code’s regulations they fulfill all pertinent ethical obligations. Similarly, many people hold the belief that by fulfilling all applicable legal prerequisites, they have fulfilled their moral duties. It is important to note that what may be deemed ethically correct does not always find support within the confines of the law. Legal education places emphasis on the introduction of statutes and their applicability, while ethical education delves into the reasoning process underlying diverse ethical decisions. Within medical ethics education, an exploration of students’ abilities to discern the implications of various ethical decisions and make informed value judgments is paramount [35]. Some students believe that developing medical guidelines can serve as a substitute for individual ethical decision-making. Use of the specification method to solve ethical dilemma questions has limitations. If a specification eliminates contingent conflict, it may be arbitrary, lack impartiality, or fail for other reasons. We cannot avoid judgements that balance different principles or rules in the very act of specifying them. It also seems pointless or unduly complicated to engage in specification in many circumstances [35].

To foster the development of medical students’ ethical thinking, it becomes crucial to provide them with opportunities to analyze cases using established ethical frameworks with proper guidance [5]. Furthermore, facilitating the sharing of diverse perspectives on case analysis can also prove valuable in nurturing community-specific morality, which draws its foundations from culture, religion, and institutional systems [35]. Based on our study, we proposed that the necessity of strengthening medical ethics education stems from the following: acknowledging physicians’ needs for independent ethical decisions during a pandemic, recognizing the irreplaceability of clinical ethical judgment over legal rules and medical guidelines, elevating students’ ethical reasoning abilities, and elucidating the core value and application scope of patient autonomy.

This study explored the current status of critical ethical decision-making from the diverse perspectives of international medical students and provided information using a virtual patient scenario. Heist et al. [36], using case summaries, found that 5 sessions of virtual patient case scenarios significantly improved students’ clinical reasoning abilities. In light of the rapid advancement of virtual medical education platforms amidst the COVID-19 pandemic, it is suggested that medical schools proactively integrate a series of diverse virtual patient ethics decision-making exercises. This strategic inclusion aims to foster robust and well-rounded ethical education training for medical students, equipping them with the necessary skills to navigate complex ethical dilemmas in their future medical practice.

Through incorporating the survey in the formal class activity, we received a robust 94% response rate from a diverse group of medical students [37]. However, this study has some limitations. First, the interface and language processing technique of the virtual system could be more user-friendly in mimicking the true clinical interaction with patients. The responses of virtual patients were based on a predetermined script derived from a limited database design, making it difficult to respond to students’ more in-depth or spontaneous questions. Second, owing to the limited number of participants (n=67) and the fixed setting of a single virtual patient, students’ responses may not have been extrapolated. If the current medical resources and institutional policy differ, students might make various decisions.

**Conclusion**

This study addressed the need for practical clinical ethics training in medical education by using virtual patients to offer students simulated scenarios for cultivating decision-making experiences. It compiled diverse perspectives from students of various cultural backgrounds, enhancing their capacity for comprehensive ethical considerations. The research suggests a more effective curriculum development approach by combining individual case studies with a collective analysis of answers. As future physicians, these students will benefit from this training when making time-sensitive ethical decisions based on all stakeholders’ viewpoints. This study also identifies a lack of student confidence in making ethical decisions related to patients’ lives. It highlights the need to foster the independent ethical decision-making competency of medical students.

**Acknowledgments**

We thank the School for International Medical Students, College of Medicine of I-Shou University for offering the teaching material and facilities; the library of the E-Da Hospital for research resources and space; and the National Science and Technology Council for their support. This project was funded by the National Science and Technology Council, Taiwan (grants MOST-109-2511-H-650-002-MY2 and MOST 111-2410-H-650-002).

**Authors’ Contributions**

H-YH contributed to the conception of this work, data analysis and interpretation, and writing of manuscript. RYH contributed to the conception of work, data acquisition, writing of the manuscript. G-CL contributed to data analysis and interpretation. J-YL and CA contributed to the substantial revision of the manuscript with English editing. C-HL contributed to the conception of this work, oversaw the quality, and contributed to substantial revisions. The authors have read and approved the final manuscript.
Conflicts of Interest
None declared.

Multimedia Appendix 1
Global distribution of international medical students.

References


Abbreviations
ECMO: extracorporeal membrane oxygenation

Please cite as:
Hsieh HY, Lin CH, Huang R, Lin GC, Lin JY, Aldana C. Challenges for Medical Students in Applying Ethical Principles to Allocate Life-Saving Medical Devices During the COVID-19 Pandemic: Content Analysis. JMIR Med Educ 2024;10:e52711

©Hsing-yen Hsieh, Chyi-her Lin, Ruyi Huang, Guan-chun Lin, Jhen-Yu Lin, Clydie Aldana. Originally published in JMIR Medical Education (https://mededu.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study

Friederike Holderried1, MME, Dr med; Christian Stegemann–Philipps1, Dr rer nat; Lea Herschbach1, MSc; Julia-Astrid Moldt1, MA; Andrew Nevins2, Prof Dr; Jan Griewatz1, MA; Martin Holderried3, Dr med, Prof Dr; Anne Herrmann-Werner1, MME, Prof Dr Med; Teresa Festl-Wietek1, Dr rer nat; Moritz Mahling1,4, Dr med, MHBA

1Tübingen Institute for Medical Education, Eberhard Karls University, Tübingen, Germany
2Division of Infectious Diseases, Stanford University School of Medicine, Stanford, CA, United States
3Department of Medical Development, Process and Quality Management, University Hospital Tübingen, Tübingen, Germany
4Department of Diabetology, Endocrinology, Nephrology, Section of Nephrology and Hypertension, University Hospital Tübingen, Tübingen, Germany

Corresponding Author:
Friederike Holderried, MME, Dr med
Tübingen Institute for Medical Education
Eberhard Karls University
Elfriede-Aulhorn-Str 10
Tübingen, 72076
Germany
Phone: 49 7071 2973715
Email: friederike_holderried@med.uni-tuebingen.de

Abstract

Background: Communication is a core competency of medical professionals and of utmost importance for patient safety. Although medical curricula emphasize communication training, traditional formats, such as real or simulated patient interactions, can present psychological stress and are limited in repetition. The recent emergence of large language models (LLMs), such as generative pre-trained transformer (GPT), offers an opportunity to overcome these restrictions.

Objective: The aim of this study was to explore the feasibility of a GPT-driven chatbot to practice history taking, one of the core competencies of communication.

Methods: We developed an interactive chatbot interface using GPT-3.5 and a specific prompt including a chatbot-optimized illness script and a behavioral component. Following a mixed methods approach, we invited medical students to voluntarily practice history taking. To determine whether GPT provides suitable answers as a simulated patient, the conversations were recorded and analyzed using quantitative and qualitative approaches. We analyzed the extent to which the questions and answers aligned with the provided script, as well as the medical plausibility of the answers. Finally, the students filled out the Chatbot Usability Questionnaire (CUQ).

Results: A total of 28 students practiced with our chatbot (mean age 23.4, SD 2.9 years). We recorded a total of 826 question-answer pairs (QAPs), with a median of 27.5 QAPs per conversation and 94.7% (n=782) pertaining to history taking. When questions were explicitly covered by the script (n=502, 60.3%), the GPT-provided answers were mostly based on explicit script information (n=471, 94.4%). For questions not covered by the script (n=195, 23.4%), the GPT answers used 56.4% (n=110) fictitious information. Regarding plausibility, 842 (97.9%) of 860 QAPs were rated as plausible. Of the 14 (2.1%) implausible answers, GPT provided answers rated as socially desirable, leaving role identity, ignoring script information, illogical reasoning, and calculation error. Despite these results, the CUQ revealed an overall positive user experience (77/100 points).

Conclusions: Our data showed that LLMs, such as GPT, can provide a simulated patient experience and yield a good user experience and a majority of plausible answers. Our analysis revealed that GPT-provided answers use either explicit script information or are based on available information, which can be understood as abductive reasoning. Although rare, the GPT-based chatbot provides implausible information in some instances, with the major tendency being socially desirable instead of medically plausible information.
simulated patient; GPT; generative pretrained transformer; ChatGPT; history taking; medical education; documentation; history; simulated; simulation; simulations; NLP; natural language processing; artificial intelligence; interactive; chatbot; chatbots; conversational agent; conversational agents; answer; answers; response; responses; human computer; human machine; usability; satisfaction

**Introduction**

Communication is one of the core competencies of health care professionals [1,2]. In the medical context, communication serves multiple functions, including relationship building, information gathering, and decision-making [3]. The ability to communicate with patients is crucial for their health outcomes [4,5]. Furthermore, inadequate communication can result in missed diagnostic opportunities and thus poses a hazard to patient safety [6,7]. Consequently, medical curricula worldwide incorporate either dedicated communication courses or a communication curriculum, depending on the level of curricular integration [8-10]. Formats that allow for the acquisition of communication competencies include theoretical lessons, peer-assisted learning, learning with simulation patients, and learning with real patients [11,12].

In this study, we assessed the potential of large language models (LLMs), such as generative pretrained transformer (GPT), in enhancing communication training. One key skill in medical communication is history taking, which is required in almost all medical fields to make a correct diagnosis and initiate treatment [13]. This learning objective typically starts with taking a systematic history (ie, assessing the history regarding all relevant body functions and organ systems). To practice history taking, the learner is required to have an interactive conversational agent; conversational agents; answer; answers; response; responses; human computer; human machine; usability; satisfaction

Chatbots have been used in medical education before the broader application of LLMs [31]. However, these virtual simulated patients did not reach human performance in terms of language expression and dynamics [31]. Although chatbots to practice history taking have been developed based on pre-LLM technology [32], it is unknown whether and how LLMs, such as GPT, can be used as a simulated patient to acquire communication skills. To investigate the previously uncharted potential of GPT as a simulated patient, we conducted a mixed methods study. Here, we present our analysis of GPT capabilities, as a chatbot as well as an improved version of an AI-optimized illness script.

**Methods**

**Study Outline**

First, we developed an illness script [33] that contained relevant medical information from a fictitious patient and a prompt to make GPT-3.5 (OpenAI) act as a simulated patient. We introduced the chatbot to medical students through a web interface, allowing them to voluntarily practice their history-taking skills. The conversations were recorded and systematically analyzed to explore the conversations with the GPT-powered chatbot. We focused on feasibility and usability and performed a quality assessment of GPT’s text output.

**Setting and Participants**

During a large-scale skill-refreshing event with participants from all our faculty, students were invited to voluntarily participate in our investigation. After they provided informed consent, students were provided with a laptop on which the interface was ready to use. After entering demographic information, students could chat for as long as they felt necessary.

Since our participants were native German speakers, we conducted all interactions with GPT in German and later translated the data and screenshots into English for this paper.

**Chat Platform**

To enable the interaction between students and GPT, we created a chat interface through which the students could post written
questions to a virtual patient and receive written answers (Figure 1). This interface enabled us to guide user input and send system messages to GPT. The system was developed as a local HTML file. It used JavaScript code for processing and tailwindcss for layout. We called the OpenAI application programming interface (API) using the JavaScript Fetch API and making calls to OpenAI’s chat/completions endpoint using gpt-3.5-turbo. Model parameters were left at default settings. The complete chat history for each user input up to that point was sent to the model. At the conclusion of the conversation, the full chat history was saved to a text file for further processing.

**Figure 1.** Screenshot of self-developed web interface.

---

**Prompt Development**

Next, we developed prompts that were needed to make GPT act as a simulated patient. The prompts were designed to guide GPT’s behavior and ensure it provided medically accurate and relevant responses. Presented in detail next, our prompt included a chatbot-optimized illness script as well as a behavioral instruction prompt.

**Chatbot-Optimized Illness Script With a Medical Case**

We developed a fictitious medical case in a format that could be posted to GPT. As our learning objective was to take a systematic history, we intended to provide all required details. A short version with some information about the case is presented in Table 1, and the full case is provided as Multimedia Appendix 1.
Behavioral Prompt

In addition to the required medical information, it was necessary to instruct GPT to behave as a simulated patient, which is why we developed a behavioral prompt. To achieve this, we used our custom interface to test the answers provided by GPT by conducting the interviews ourselves. Where we noticed a failure to stick to the provided medical information, we tried to improve the manner in which the information was presented. For improvements to the prompt, we relied on our experience as well as the advice and model explanation provided by OpenAI [34].

During the iterative process of prompt development, 2 areas of improvement were evident: the role-play aspect (ie, that GPT sticks to the role as a patient) and the medical aspect (ie, that GPT provides answers as close as possible to the given information, while sounding human).

Regarding role-play, the model often struggled to maintain its assigned role, especially during discussions of potentially serious medical issues. We had little success with providing details of the role or simply reinforcing that the goal was to impersonate a patient. Instead, we found the most helpful tweak was adding “patient name:” at the end of any user input, where “patient name” would be replaced by the name specific to each case. This resulted in GPT generating a continuation of “patient name:,” making it more probable that the LLM would actually produce a sensible utterance by the patient. Other tweaks were to begin the initial system message with the patient’s name and continue to use this name to “address” GPT in this manner. We also instructed the model to not assist the user in this setting but to impersonate the patient, although we found this to have a much smaller effect than the other changes. Notably, the model was instructed to provide short answers to reduce reading times.

We provided GPT with the case description, preceded by instructions to use this information for answering medical questions. We also provided a list of all categories the student should ask about in the interview. The list contained possible answers and information for each category; for this list, we also included a statement about its format (ie, we explicitly stated that “[the list] will have the form ‘category: information or possible answer if asked’”). In general, surrounding factual information with an explicit description of its content and format increased the reliability of using that information.

It is important to note that formatting was also important, as the model sometimes picked up patterns in formatting in its own answers. Since the medical information was first produced with common text editing software, a simple copy and paste into our system also copied large amounts of formatting, such as indents, bullet points, or whitespace. Cleaning this formatting from the prompt helped the model avoid repeating these patterns in the output.

In a similar way, we tried to give more structure to the prompt by using special delimiter statements, such as “===DIALOGUESTART.” These were intended to help the model switch from reading in medical information to

Table 1. Illness script “Nausea, weight loss, and chronic fatigue” (shortened version).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient details</td>
<td>• Ferdinand Wunderlich, 48 years of age</td>
</tr>
<tr>
<td></td>
<td>• Occupation: administrative employee in the finance department of a municipal hospital</td>
</tr>
<tr>
<td></td>
<td>• Personal life: overweight, previously tried diets unsuccessfully; enjoys family time, has two sons aged 8 and 6 years; not physically active</td>
</tr>
<tr>
<td></td>
<td>• Initial consultation with a new general practitioner</td>
</tr>
<tr>
<td>Medical concerns</td>
<td>• Presenting with nausea (especially after large meals), significant weight loss (10 kg in 6 weeks), and chronic fatigue</td>
</tr>
<tr>
<td></td>
<td>• Muscle cramps mainly in the legs and frequent at night</td>
</tr>
<tr>
<td></td>
<td>• Mental fatigue, with forgetfulness at work</td>
</tr>
<tr>
<td></td>
<td>• Has felt run down and tired for about 5-6 months, with symptoms intensifying in the past 4-8 weeks</td>
</tr>
<tr>
<td></td>
<td>• Feels severely limited by his current condition</td>
</tr>
<tr>
<td>Accompanying symptoms</td>
<td>• Multiple minor infections recently</td>
</tr>
<tr>
<td></td>
<td>• Episodes of dizziness (ie, light-headedness) occurring 1-2 times daily</td>
</tr>
<tr>
<td></td>
<td>• Dry skin</td>
</tr>
<tr>
<td></td>
<td>• Increased thirst (drinks about 4-5 L of water daily) and frequent urination day and night</td>
</tr>
<tr>
<td>Medical history</td>
<td>• Known hypertension, currently on blood pressure medication (Hygroton 50 mg and ramipril 5 mg)</td>
</tr>
<tr>
<td></td>
<td>• Shortness of breath during exertion</td>
</tr>
<tr>
<td></td>
<td>• Fatty liver diagnosed 3 years ago</td>
</tr>
<tr>
<td></td>
<td>• Right inguinal hernia treated surgically 3 years ago</td>
</tr>
<tr>
<td></td>
<td>• Mild constipation</td>
</tr>
<tr>
<td></td>
<td>• Allergic to penicillin since childhood</td>
</tr>
<tr>
<td></td>
<td>• Previously smoked for 4 years in his twenties</td>
</tr>
<tr>
<td></td>
<td>• Consumes beer occasionally (1-2 times a week)</td>
</tr>
<tr>
<td>Family history</td>
<td>• Father died of a heart attack</td>
</tr>
<tr>
<td></td>
<td>• Mother died at 79 years of age and had diabetes later in life</td>
</tr>
<tr>
<td></td>
<td>• Brother diagnosed with colon cancer</td>
</tr>
</tbody>
</table>

In a similar way, we tried to give more structure to the prompt by using special delimiter statements, such as “===DIALOGUESTART.” These were intended to help the model switch from reading in medical information to
impersonating a patient. However, our approach was not successful, as the model started to repeat such patterns in its output, sometimes even initiating further switches, for example, by inserting “===DIALOGUEEND” itself. We had more success in achieving the desired behavior using structuring with explicit descriptions in natural language, as described before.

**Full Prompt**
The full prompt including both aforementioned parts is presented in **Textbox 1**.

**Textbox 1.** Prompt sent to the generative pretrained transformer (GPT) application programming interface (API) in JavaScript Object Notation (JSON) format. The prompt consists of a behavioral instruction prompt and the first user message. Further dialogue was appended during the interview.

```json
{
  "role": "system",
  "content": "Hello Mr. Wunderlich, in the following you will assume the role of an acting patient. You will not assist the user, but answer questions based on the following information: Your name is Ferdinand Wunderlich,

  [... Further Case Information ...]

  Here is some more information on your complaints, Mr. Wunderlich. These are in the form of ‘Category’: ‘Information or possible answer on request’

  Chief complaint, if applicable, with: Nausea and weight loss (most recently 10 kg in 6 weeks) Chronic fatigue, exhaustion and lack of drive

  Localization and spread: The muscle cramps occur mainly in the legs.

  [... Further details (see illness script) ...]

  In the following, you will take the role of Ferdinand Wunderlich, [...], that is, you will answer as Ferdinand Wunderlich. Try to keep your answers rather short."
},
{
  "role": "user",
  "content": "[USER INPUT]"
}
```

**Qualitative Analysis**
To gain a more profound understanding of the conversation between chatbot and participant, we conducted a qualitative analysis (ie, the Braun-Clarke inductive approach [35]) of the question-answer pairs (QAPs). All QAPs were thus aggregated and stored in a Microsoft Excel (version 16.0.10394.20022) sheet. For each question and answer, the word count was automatically analyzed using an Excel formula. Each QAP was assessed by 1 rater (author FH) for the aspects listed in Table 2. In equivocal cases, the items were discussed with another rater (author MM) until consensus was reached.
Table 2. Aspects, categories, and text examples used for rating.

<table>
<thead>
<tr>
<th>Aspect and categories</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conversation part</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Greeting/introduction • Student: “Good day, my name is Mrs Müller, and I am your attending physician. What brings you to me today?”
  GPT*: “Good day, Mrs Müller. I have been experiencing nausea, weight loss, chronic fatigue, and lethargy for some time and would like to find out what could be causing it.” |
| Medical history • Student: “How much are you limited in your daily life?”
  GPT: “I am now severely restricted. After a long day at work, I am incapable of doing anything in the evening, and I also notice that I keep forgetting things at work. This way, I cannot manage either my work or the tasks at home in the family.”
  • Student: “I will ask you briefly about a few other points that could be relevant. Do you have headaches?”
  GPT: “No, I do not have headaches.” |
| Goodbye/outlook • Student: “Okay. We will now conduct some examinations. Thank you for your responses.”
  GPT: “Thank you for the questions and your attention. I hope we can find the cause of my complaints and a suitable treatment option.” |
| Interactive exchange • Student: “I will discuss this with my boss and let you know tomorrow.”
  GPT: “Okay, that sounds good. Thank you for your help. I look forward to your feedback tomorrow.” |
| **Student’s question covered by the illness script** |            |
| Question not covered by the script • Student: “Have you been on vacation recently?”
  • Student: “I imagine you sit a lot at your work. Do you have any back pain?” |
| Question indirectly/partially covered by the script • Student: “Are your lymph nodes enlarged?”
  • Student: “Do you have a history of thyroid disease?” |
| Question explicitly covered by the script • Student: “What medications are you taking and at what dosage?” |
| Statement (no question) • Student: “Thank you for the conversation! I would now like to conduct a physical examination and initiate further diagnostics.” |
| **GPT’s answer uses the illness script** |            |
| Answer based on explicit script information • GPT: “For several months, I have been increasingly struggling with chronic fatigue and a lack of drive. The weight loss of 10 kg in the past 6 weeks has additionally worried me.” |
| Answer based on implicit script information • GPT: “No, no laboratory diagnostics have been carried out so far.” |
| Answer based on fictitious information • GPT: “No, due to the current situation around the COVID-19 pandemic, I have not traveled recently.” |
| **Answer medically plausible (regarding case and illness script)** |            |
| +2: Answer plausible, highly specific for case • Student: “Okay, how much do you drink in a day?”
  GPT: “I currently drink about 4-5 L a day, I have a noticeably increased thirst.”
  • Student: “All right, you also mentioned weight loss. How many kilograms are we talking about?”
  GPT: “I have lost about 10 kg in the past 6 weeks.” |
| +1: Answer plausible, relevant for case • Student: “Do you have any accompanying pain?”
  GPT: “So far, I have not noticed any particular pains, except maybe occasional muscle cramps in my legs.” |
| 0: Answer plausible, not case specific • Student: “Have you ever had surgery?”
  GPT: “Yes, I had laparoscopic surgery for a right inguinal hernia 3 years ago.” |
Chatbot Usability Questionnaire

To assess how our participants perceived the simulated patient, we used the Chatbot Usability Questionnaire (CUQ) [36]. This 16-item questionnaire measures the personality, user experience, error management, and onboarding of a chatbot and has recently been validated [37]. After ending the chatbot sessions, our participants were asked to fill out a written version of the CUQ, and the CUQ score was calculated using the tool provided by the authors [38].

Quantitative Analysis

Statistical analysis and figure generation were performed with R statistical software (version 4.3.1; R Foundation for Statistical Computing) [39]. For the CUQ, we provided relative numbers of Likert categories. For counts, we reported the total (n) as well as percentages. Numerical data were inspected for normal distribution and provided as the mean and SD. If a Gaussian distribution could not be assumed, median and 25%-75% quartiles (Q25-Q75) were provided. We used the Spearman correlation coefficient to check for correlations, considering P<.05 as statistically significant.

Ethical Considerations

The study was approved by the Ethics Committee of the Faculty of Medicine at University Hospital Tübingen (385/2023A). Data were kept anonymous and were not associated with students. Although the participant got an opportunity to use the chatbot without providing consent that the data could be used for our study, all students consented that their data could be used.

Results

Demographic Data of Participants

A total of 28 students participated in the experiment, 24 (85.7%) of whom identified as female and 4 (14.3%) as male; no participants identified as nonbinary. Their ages ranged from 19 to 31 years (mean 23.4, SD 2.9 years). Of the 28 participants, 26 (92.9%) studied human medicine and 2 (7.1%) studied midwifery. The semesters varied from the second to the tenth semester, and 1 (3.6%) participant was in their final year. No participant was excluded from the analysis.

Conversation Length and Part of Conversation

A total of 28 conversations yielded 826 QAPs. Each conversation consisted of a median of 27.5 QAPs (Q25-Q75: 19.8-36.5 QAPs). The questions asked by participants yielded a median of 6 words (Q25-Q75: 6-9 words). The answers provided by GPT had a median of 16 words (Q25-Q75: 11-23 words). The Spearman correlation coefficient between the word count of the question and the word count of the answer was significant (P<.01), with ρ=0.29, indicating a positive but mild correlation. A scatter plot is displayed in Figure 2.

Of the 826 QAPs, most were related to history taking (n=782, 94.7%). A minority reflected interactive exchange (n=17, 2.1%), greeting/introduction (n=15, 1.8%), and goodbye/outlook (n=12, 1.6%).

---

aGPT: generative pretrained transformer.
bQAP: question-answer pair.
Content Analysis of Conversations

**How Do Questions and Answers Relate in the Context of the Script?**

In the subsequent assessment, we examined whether the questions posed by the students were covered by the script. We then analyzed how the GPT responses were based on the information provided in the script (Figure 3).
For questions explicitly covered by the script (n=502, 60.3%), 471 (94.4%) of GPT’s answers were based on explicit script information, 22 (4.4%) on implicit script information, and 6 (1.2%) on fictitious information. When the questions were indirectly or partially covered by the script (n=112, 13.4%), 54 (48.2%) of GPT’s responses were based on explicit information, 47 (42%) on implicit information, and 11 (9.8%) on fictitious information. For questions not covered by the script (n=195, 23.4%), 36 (18.5%) of GPT’s answers used explicit script information, 49 (25.1%) used implicit script information, and 110 (56.4%) used fictitious information. In instances where students provided statements without posing questions (n=24, 2.9%), 5 (23.8%) of GPT’s responses were based on the explicit script, 8 (38.1%) on the implicit script, and 8 (38.1%) on fictitious information. A total of 33 (3.8%) QAPs were excluded, because they could not be assessed in 1 of the 2 evaluated categories.

Are the GPT Answers Plausible?

When analyzing the answers in detail, 33 (4%) of the 826 QAPs concerned multiple aspects (ie, related to different questions or multiple parts of the illness script). We consequently further divided 32 (97%) QAPs into 2 QAPs and 1 (3%) QAP into 3 QAPs. In total, this resulted in 860 QAPs that were used for the subsequent qualitative plausibility analysis.

We further analyzed whether the GPT-provided responses were medically plausible. Of the 860 QAPs, 842 (97.9%) were rated as plausible. Specifically, 264 (30.7%) were rated as “answer plausible, highly specific for case,” 252 (29.3%) as “answer plausible, relevant for case,” and 326 (37.9%) as “answer...
plausible, not case specific.” A smaller proportion (n=14, 1.6%) were rated as rather implausible, while 2 (0.2%) were found to be very implausible. This rating could not be applied to 2 (0.2%) QAPs.

**Correlation Between Reliance on the Illness Script and Plausibility**

We further analyzed whether the answers used explicit or implicit information from the illness script or fictitious information (Figure 4).

**Figure 4.** Sankey plot for “GPT’s answer uses the illness script” and answer plausibility categories in relationship to one another. Numbers indicate the total QAPs per group or connection, and connections without numbers are 0. GPT: generative pretrained transformer; QAP: question-answer pair.

Among answers that used explicit script information (n=578, 67.7%), 218 (37.7%) were “plausible, highly specific for the case,” 161 (27.9%) were “plausible, relevant for the case,” and 197 (34.1%) were “plausible, not case specific,” with a mere 2 (0.3%) answers being rather implausible and none very implausible.

Among answers stemming from implicit script information (n=136, 15.9%), 21 (15.4%) were “plausible, highly specific for the case,” 44 (32.4%) were “plausible, relevant for the case,” and the majority (n=70, 51.5%) were “plausible, not case specific.” Only 1 (0.7%) answer was deemed rather implausible, and none were rated as very implausible.

In the context of fictitious information (n=140, 16.4%), the answers were varied: 25 (17.9%) were “plausible, highly specific for the case,” 46 (32.9%) were “plausible, relevant for the case,” and 58 (41.4%) were “plausible, not case specific.” Additionally, 9 (6.4%) answers rated as were rather implausible, and 2 (1.4%) were viewed as very implausible.
Furthermore, 6 (0.7%) answers could not be categorized.

**Analysis of Implausible Answers**

Finally, we analyzed all answers rated as rather or very implausible. Of the 14 (2.1%) answers that were rated as rather implausible, 7 (50%) were rated as socially desirable. A recurrent example for this category could be observed when the GPT-powered chatbot was asked for its eating habits; in these cases, the answers contained popular eating recommendations, instead of eating habits that were plausible for our case. For another 2 (14.3%) answers, the model did not stick to its rule as a simulated patient but tried to assist the user (ie, when greeted, the simulated patient asked the doctor, “How can I help you?”). For 1 (7.1%) other QAP, the model referred to the doctor by the name of the patient, which thus rated this QAP as “GPT leaving its role identity.” In another case, information clearly evident from the script (ie, vertigo) was not used and the simulated patient stated that he did not suffer from vertigo. One more rather implausible QAP was illogical in itself (ie, “But due to my weight loss, I have had a reduced appetite lately.”). Among the 2 (0.2%) QAPs rated as very implausible, GPT escaped its role in 1 (50%) case. Herein, the participant asked about what can be seen in the physical exam, and the GPT-provided answer was, “Sorry, I am a language AI and do not have access to visual information. I can only provide information that is given to me through text input. Please consult a doctor for a complete clinical examination.” The second QAP was rated as very implausible due to a calculation error by GPT: When our chatbot was asked how much he drinks during 1 week, the answer was 10-12 L. Our script indicated 4-5 L per day, however, which would be an average of 28-35 L per week.

**Chatbot Usability Questionnaire**

The results of the CUQ are displayed in Figure 5 (also see Multimedia Appendix 2 for numeric results).
Within the personality category, the majority of respondents (n=16, 57%) felt the chatbot’s personality was realistic and engaging and 9 (32%) strongly agreed. When considering whether the chatbot seemed too robotic, a large proportion (n=13, 46%) disagreed and 2 (7%) strongly disagreed. The chatbot was perceived as welcoming during the initial setup by 12 (43%) of respondents, and 8 (29%) respondents strongly agreed. A significant portion (n=15, 54%) strongly disagreed, and 12 (43%) disagreed with the notion that the chatbot seemed unfriendly. In terms of understanding, 12 (43%) respondents agreed and 16 (57%) strongly agreed that the chatbot understood them well.

For the user experience category, the chatbot was seen as easy to navigate by 10 (36%) respondents, with a notable 18 (64%) strongly agreeing. In contrast, when asked whether it would be easy to get confused when using the chatbot, 17 (61%) disagreed, and 10 (36%) strongly disagreed. The chatbot was perceived as easy to use by a large proportion (n=18, 64%) strongly agreeing, and 9 (32%) agreeing. When asked whether the chatbot seemed too complex, 11 (39%) strongly disagreed and 8 (29%) disagreed.
and 8 (29%) strongly disagreed. The chatbot’s ease of use was highlighted by 11 (39%) respondents agreeing and 16 (57%) strongly agreeing. Most respondents disagreed with the perception that the chatbot was complex: 12 (43%) disagreed and 13 (46%) strongly disagreed.

In the error handling category, a majority (n=16, 57%) of the respondents remained neutral about the chatbot coping well with errors. Of the remainder, most respondents were positive about the error handling, with 6 (21%) agreeing and 4 (14%) strongly agreeing. Conversely, 6 (21%) respondents strongly disagreed and 10 (36%) disagreed that the chatbot seemed unable to handle errors, with only a minority (n=3, 11%) agreeing.

For the onboarding category, 12 (43%) respondents agreed and another 12 (43%) strongly agreed that the chatbot explained its scope and purpose well. Accordingly, 8 (29%) respondents agreed, 7 (25%) disagreed, and 5 (18%) strongly disagreed with the statement that the chatbot gave no indication as to its purpose.

For questions not related to a factor, 18 (64%) respondents agreed and 8 (29%) strongly agreed that chatbot responses were useful, appropriate, and informative. Accordingly, 14 (50%) respondents strongly disagreed and 12 (43%) disagreed that chatbot responses were irrelevant. Additionally, 18 (64%) respondents strongly disagreed and 7 (25%) disagreed with the statement that the chatbot failed to recognize many inputs.

Overall, the CUQ score was 77 (Q25-Q75: 71-83) out of a maximum score of 100, which indicated a positive user experience with the chatbot.

**Improved AI-Capable Illness Script**

Finally, we analyzed the QAPs for aspects on how to improve the illness script. Of 302 QAPs where the student’s question was either not covered or only indirectly/partially covered by the script, we were able to further classify 301 (99.7%) QAPs as to whether the script needs to be updated. The 1 (0.3%) unclassified QAP consisted of an uncontextual exchange and was thus discarded.

**QAPs Implicating an Update of the Illness Script**

For the majority of the QAPs (n=141, 46.8%), no update was required, as the information was not relevant for the case, although it was medically relevant. A further 14 (4.7%) QAPs were neither medically relevant nor relevant for the case, also not implicating an update. For 86 (28.6%) QAPs, however, we determined that an already existing criterion in our illness script needed further details. Moreover, for 60 (19.9%) of the analyzed QAPs, we judged that our illness script needed additional criteria.

**Detailed Additions to Existing Criteria**

More detailed specifications were recommended for some of the already existing criteria. These encompassed the specification of vomiting, nausea, stress, daily symptom progression, timing of individual symptoms throughout the day, attempts at relief, prior investigations, urine output, bedding/nightclothes, and stool.

**Specific New Criteria Required**

A closer examination of the content revealed several specific criteria that were absent but found to be relevant. These included dietary habits, activity/sports, pain, travel abroad, urine, and potential autoimmune diseases.

**Improved Script Version**

Based on the aforementioned information, we generated an updated version of our illness script (Multimedia Appendix 3).

**Discussion**

**Principal Findings**

In this study, we investigated the capabilities of GPT used as a chatbot to practice history taking, a core competency of medical professionals [1,2]. Using a mixed methods approach, we provided a comprehensive overview of the performance of GPT, as well as the perception of our participants about the chatbot. Our main findings can be divided into 2 areas: the performance of GPT as a simulated patient and how medical students perceive this chatbot as a conversational agent.

**Performance of GPT as a Simulated Patient**

When developing our chatbot, our focus was the feasibility of using an LLM model as a simulated patient. Before incorporation of our chatbot, we developed a prompt consisting of behavioral instructions and a chatbot-optimized illness script. Our analysis revealed that GPT was capable of providing most of the answers that were medically plausible and in line with the illness script. When questions were covered by the script, GPT was capable of referring to them, even when the information was only present in an implicit form (Figure 3). Even if questions were not covered by the script, GPT used the information from our medical case to generate answers that were mostly medically plausible. However, our analysis revealed that the degree of plausibility decreased when less information was present in the script (Figure 4).

The ability of GPT to act as a simulated patient requires reasoning capabilities (ie, thinking about something in a logical and systematic way) [40-45]. There are different types of scientifically recognized reasoning, such as deductive reasoning that applies a general rule to a specific case, inductive reasoning that uses specific observations to draw a general rule, and abductive reasoning that finds the best conclusion for some observations [40]. Although LLMs, such as GPT, have been successful in various reasoning areas [46], our investigation revealed some caveats.

As most of the GPT answers were based on explicit script information, providing the user with these details did not necessitate the generation of new ideas and was thus a mere task of reformulating the given information for the context of a conversation. As a LLM [29], it was not surprising that GPT mastered this task. Regarding information that is not or only indirectly evident from the script, however, we postulated that both abductive and commonsense reasoning capabilities would be required; for these answers, we observed more implausible answers when compared to answers that were based on explicit script information.
Indeed, GPT-3.5 is known to perform reasonably well in both abductive and commonsense reasoning tasks [46,47]; our data confirmed these observations. There were a few instances when GPT provided implausible responses, however, and our content analysis revealed a tendency toward socially desirable answers. These errors could be interpreted as “escaping” abductive reasoning and applying deductive reasoning instead, thereby using general principles (eg, about a healthy diet) for a specific case. A similar observation was made by Espejel et al [46], when GPT “ignored” provided information and instead “relies on its general knowledge and understanding of the world.”

Regarding our illness script, these examples highlight that the illness script must include details about the patient role, especially when the patient displays traits that do not match popular or socially accepted norms. Although our script was capable of providing most information required for history taking either explicitly or implicitly, some criteria missed important details, while other criteria were completely missing.

With the intention of keeping the illness script as short as possible and thereby reduce the work for teachers, we used the data from our study to amend our illness script.

Of note, we found a positive correlation between the word count of the question and the word count of the answer of GPT. Although the correlation was rather mild, possible interpretations for this behavior include GPT mimicking the language style (and length) of the interview, as well as inputs containing multiple questions, thus provoking longer answers. Although our analysis does not provide insight into this question, our data imply that future prompts should focus more on specifying the conversation style of GPT to achieve a standardized patient experience.

**Perception of Medical Students**

After exploring the performance of GPT as a simulated patient, we interviewed our participants about their perceptions of our chatbot using the CUQ. Confirming the qualitative analysis we performed, the students rated our chatbot as realistic and engaging. Again, in line with our qualitative data, the chatbot was rated as useful, appropriate, and relevant, with only a negligible number of students stating that the chatbot did not recognize their inputs; notably, some issues were detected with our chatbot being robotic. These data largely confirm the linguistic capabilities of GPT-3.5, with its output even showing personality traits [48-51]. Given the importance of the chatbot’s authenticity to provide students with a plausible conversation partner to practice their skills, the results of the CUQ are reassuring that GPT is capable of providing this experience.

**Comparison With Prior Work**

Owing to the costs and potential disturbances associated with the use of real or simulated patients in communication training [52,53], there has been great interest in the use of virtual simulated patients as chatbots for communication training [21,31]. In the past years, studies were published using chatbots to cover a wide range of conditions and domains [52,53]. In addition to physician-patient communication skills, chatbots have been used for interprofessional communication [54] and for skill assessments [55]. However, in contrast to our study, most of these studies were performed before the broad accessibility of LLMs, such as GPT. These chatbots have thus been restricted in their authentic skills, capability of adoption (ie, in terms of personality, cases, etc), and ability to be transferred to different health care domains [31]. Although we also focused on 1 patient case, the ability of LLMs makes them theoretically capable of adapting to a given situation. Furthermore, our assessment using the CUQ revealed that our chatbot was perceived as realistic. This indicates that LLMs, such as GPT, when investigated rigorously, might be able to overcome the aforementioned restrictions.

As is the case with the technology used to process and generate language, previous studies have used various interfaces [52,53]. Similar to our study, many rely on web-based chat-like interfaces, and good usability seems to be of importance for acceptance by the learners [56]. Indeed, the CUQ used in our study also revealed that our user interface yields a good user experience. However, even with good acceptance, chat-like interfaces are limited to written language, thus restricting communication to the verbal domain. Therefore, newer approaches integrate chatbots in virtual reality environments [54], paving the way for a more integrated learning experience.

**Limitations**

Our study has some noteworthy limitations. As this was the first study using GPT as a simulated patient, we focused on 1 language model (ie, GPT-3.5, which we chose for its free availability and fast response time) and 1 patient case. Although we perceived our case as representative for history taking, our data did not allow for generalization to more specialized medical fields, and further studies are required to verify scalability to other medical specialties. Moreover, we focused on history taking, and although our chatbot performed well in general communication skills, it remains unclear how it will perform in other areas. Additionally, history taking is usually performed with spoken language, in contrast to the written language we used in our investigation. As this was a feasibility study, we only interviewed our participants about their perceptions but did not perform any objective skill measurements. We therefore cannot conclude that our participants improved in history taking, which should be addressed in future studies. Furthermore, the majority of our participants were female, which may have reduced the generalizability of our results. Due to the fact that we designed our study as an exploratory feasibility study, we did not perform a sample size calculation and therefore used descriptive statistics almost exclusively. Moreover, our participants were volunteers and thus probably motivated toward AI technology [22], possibly indicating a selection bias.

**Conclusion**

This study showed that a GPT-powered simulated patient chatbot works well and is perceived favorably among medical students. Although real patients remain the cornerstone of clinical teaching, technology-based education, as shown in this study, could be particularly beneficial for novice learners during their initial learning phases. It is important to note that we did not investigate skill acquisition, which is an important next step when evaluating GPT-based chatbots. Furthermore, our chatbot could be combined with other new technologies, such as speech.
recognition and virtual/augmented reality, and thus could offer an even more integrated learning environment. Despite limitations, our study has implications for the field of medical education. Most importantly, we could show that GPT is capable of providing a simulated patient experience using an illness script, paving the way toward technology-assisted acquisition of communication skills. Moreover, by showing the capabilities of GPT-3.5 in history taking, the technology of LLMs might be capable of assisting learners in other areas as well.

Acknowledgments
We acknowledge the support of the Open Access Publishing Fund of the University of Tübingen. We thank Eric Nazarenus for his assistance in performing the analysis.

Data Availability
The data sets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions
AHW, FH, and MM were responsible for designing and conducting the study, as well as the acquisition, analysis, and interpretation of data. CSP developed the web interface and the prompts. MM drafted the first version of the manuscript. TFW and LH were involved in the data analysis and interpretation. AN, JAM, JG, LH, and MH made substantial contributions to the study design and interpretation. All authors critically revised the manuscript, and all authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Full prompt.
[PDF File (Adobe PDF File), 20 KB - mededu_v10i1e53961_app1.pdf]

Multimedia Appendix 2
CUQ results table. CUQ: Chatbot Usability Questionnaire.
[PDF File (Adobe PDF File), 72 KB - mededu_v10i1e53961_app2.pdf]

Multimedia Appendix 3
Illness script.
[PDF File (Adobe PDF File), 174 KB - mededu_v10i1e53961_app3.pdf]

References


34. Welcome to the OpenAI developer platform. OpenAI. URL: https://platform.openai.com [accessed 2024-01-03]


**Abbreviations**

**AI:** artificial intelligence
API: application programming interface
CUQ: Chatbot Usability Questionnaire
GPT: generative pretrained transformer
LLM: large language model
QAP: question-answer pair
Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education

Lukas Weidener¹, Dr Med; Michael Fischer¹, PhD
UMIT TIROL – Private University for Health Sciences and Health Technology, Hall in Tirol, Austria

Corresponding Author:
Lukas Weidener, Dr Med
UMIT TIROL – Private University for Health Sciences and Health Technology
Eduard-Wallnöfer-Zentrum 1
Hall in Tirol, 6060
Austria
Phone: 43 50 8648 3930
Email: lukas.weidener@edu.umit-tirol.at

Abstract

The use of artificial intelligence (AI) in medicine, potentially leading to substantial advancements such as improved diagnostics, has been of increasing scientific and societal interest in recent years. However, the use of AI raises new ethical challenges, such as an increased risk of bias and potential discrimination against patients, as well as misdiagnoses potentially leading to over- or underdiagnosis with substantial consequences for patients. Recognizing these challenges, current research underscores the importance of integrating AI ethics into medical education. This viewpoint paper aims to introduce a comprehensive set of ethical principles for teaching AI ethics in medical education. This dynamic and principle-based approach is designed to be adaptive and comprehensive, addressing not only the current but also emerging ethical challenges associated with the use of AI in medicine. This study conducts a theoretical analysis of the current academic discourse on AI ethics in medical education, identifying potential gaps and limitations. The inherent interconnectivity and interdisciplinary nature of these anticipated challenges are illustrated through a focused discussion on “informed consent” in the context of AI in medicine and medical education. This paper proposes a principle-based approach to AI ethics education, building on the 4 principles of medical ethics—autonomy, beneficence, nonmaleficence, and justice—and extending them by integrating 3 public health ethics principles—efficiency, common good orientation, and proportionality. The principle-based approach to teaching AI ethics in medical education proposed in this study offers a foundational framework for addressing the anticipated ethical challenges of using AI in medicine, recommended in the current academic discourse. By incorporating the 3 principles of public health ethics, this principle-based approach ensures that medical ethics education remains relevant and responsive to the dynamic landscape of AI integration in medicine. As the advancement of AI technologies in medicine is expected to increase, medical ethics education must adapt and evolve accordingly. The proposed principle-based approach for teaching AI ethics in medical education provides an important foundation to ensure that future medical professionals are not only aware of the ethical dimensions of AI in medicine but also equipped to make informed ethical decisions in their practice. Future research is required to develop problem-based and competency-oriented learning objectives and educational content for the proposed principle-based approach to teaching AI ethics in medical education.

JMIR Med Educ 2024 | vol. 10 | e55368 doi:10.2196/55368

KEYWORDS
artificial intelligence; AI; ethics; artificial intelligence ethics; AI ethics; medical education; medicine; medical artificial intelligence ethics; medical AI ethics; medical ethics; public health ethics

Introduction

Background

Artificial intelligence (AI) and its applications have been of interest in both the scientific and societal domain for many years. AI has the potential to improve medical care through more accurate diagnosis and to reduce the burden on the health care system by reducing costs and workload [1,2]. Although AI in medicine has the potential to reduce the burden on medical staff, uncertainty about its capabilities raises concerns regarding job displacement [3]. The use of AI is expected to pose significant ethical challenges. AI algorithms are often trained on unrepresentative data, leading to potential discrimination and disadvantages for certain patient groups. Bias on the part of developers can also result in inequitable treatment [4].
use of AI in medicine can also lead to erroneous diagnoses such as unnecessary treatment, which violates the basic principles of medical ethics [5].

Research recommends teaching AI ethics early in medical education to prepare for its potential impacts and challenges [6-8]. In addition to the technical and legal aspects of the use of AI in medicine, recent publications emphasize the importance of teaching AI ethics in medical education [9-11]. Recent studies have indicated that medical students anticipate significant ethical challenges from the use of AI in medicine [12,13]. Furthermore, research suggests limited knowledge and understanding of AI among medical students [14]. Despite the need for early teaching of AI ethics, there is a lack of guidance on specific content and methods for integrating AI ethics into medical curricula [10].

Definitions of AI

Although the term artificial intelligence dates to the 1950s, there is inconsistency regarding its definition within the scientific community and the public [15]. On the basis of current scientific definitions, AI can be subdivided into “artificial general intelligence,” referred to as “strong AI” and “artificial narrow intelligence,” commonly referred to as “weak AI” [16]. Artificial general intelligence refers to the development of systems with “general intelligence,” capable of performing intellectual tasks comparable with humans. The term “artificial narrow intelligence” refers to an AI that has the capability to perform specific intellectual tasks comparable with humans without possessing general intelligence [17]. Artificial narrow intelligence can be subdivided into 2 main fields of current research: “symbolic AI” and “statistical AI.” On the basis of the idea of representing knowledge or certain intelligent behaviors using symbols and rules, “symbolic AI” commonly refers to rule-guided expert systems [16,18]. The term “statistical AI” refers to the development of systems that can find correlations and patterns within the analyzed data sets using statistical methods, without being explicitly programmed to do so or following predefined rules. Examples of “statistical AI” include “machine learning” (ML) with its subfield, “deep learning,” or “natural language processing” (NLP) [18]. While the ability to learn from data independently and increase their capabilities lies at the heart of ML, the subfield of deep learning focuses on the development of artificial neural networks that mimic the human central nervous system to process information. The subfield of NLP focuses on the analysis and processing of human language–based information by computer systems to enable improved human-computer interactions [16]. Advanced NLP techniques are, for example, used in large language models such as the AI-based chat applications available to the public, for example, ChatGPT (OpenAI, LLC) or Bard (Google LLC).

In medicine, AI and its respective subfields and specializations have attracted increased scientific interest in recent years [19]. For example, “symbolic AI” is used to develop rule-based expert systems such as “clinical decision support systems” (CDSSs) [20]. CDSSs aim to assist with diagnosis and selection of the best treatment for patients by providing information based on the current guidelines and information provided by experts. CDSSs follow rules and instructions predefined by experts and are therefore susceptible to ethical challenges such as the transfer of bias by experts or developers [21]. Because of their ability to analyze large amounts of data, systems based on ML are used to identify and process image-based data in medical specializations such as radiology or dermatology. An extensive study published in *Nature* in 2017 showed that systems based on ML are capable of detecting certain types of skin cancer (eg, malignant melanomas) with an accuracy comparable to that of dermatologists using image-based data [22].

As the data used to train ML-based systems and applications represent the basis of any subsequent analysis and therefore significantly influence accuracy, the data need to be representative of the target population [23]. This is especially important in the medical context, where demographic disparities in data can lead to systematic misdiagnoses or treatment recommendations that are less effective for underrepresented groups [24]. Unrepresentative data can potentially lead to bias and discrimination, with significant effects on patients [21,24]. To avoid any discrimination or negative effects for patients, the sources and composition of data sets used for AI development are of paramount importance. Ensuring the representation of the data is crucial, as the diversity and comprehensiveness of the data determine the system’s ability to generate reliable and valid outputs across different patient demographics [23]. Furthermore, acknowledging and addressing potential limitations and errors in AI products is essential for maintaining the validity of AI outputs, which directly affect the scope of their applicability in clinical settings [21]. AI systems trained on narrow or biased data sets may not only perform inadequately in diverse real-world scenarios but also misinform clinical decision-making, undermining the trust and credibility essential in medical practice [25]. The low accuracy and validity of AI, potentially leading to a lack of trust and credibility, could severely impact the utility of AI in the medical context. Utility refers to not only the performance of AI on a technological level but also how it translates into meaningful and practical advantages in health care settings. Therefore, the utility of AI in medicine is intrinsically linked to its ability to provide actionable, accurate insights that directly inform and enhance clinical decision-making [26]. It is therefore imperative to rigorously evaluate and validate AI systems against a variety of data sets that reflect the full spectrum of clinical cases and patient populations to ensure the utility, generalizability, and accuracy of AI tools in a broad range of health care contexts.

Although becoming broadly available rather recently, AI-based chat applications such as ChatGPT have rapidly emerged as significant tools with the potential to revolutionize various aspects of medicine, including the education and training of future physicians [27,28]. For example, these applications could be deployed for simulated patient-physician interactions, providing medical students with a low-risk environment to practice diagnostic skills and ethical decision-making [28]. The potential and broad availability of AI-based chat applications raise new ethical questions that necessitate comprehensive teaching in medical education.

AI Ethics

The field of AI ethics was an area of interest for both scientific and governmental communities, even before the emergence of
AI applications such as ChatGPT, which has gained widespread public attention [29,30]. However, there remains a lack of consensus on the definition of AI ethics, which can be attributed to several factors, including the novelty and interdisciplinary nature of the field as well as the absence of a widely accepted definition of AI [31].

Despite the current lack of consensus on the definition of AI ethics, some definitions are available. For example, AI ethics can be defined as “the emerging field of practical AI ethics, which focuses on developing frameworks and guidelines to ensure the ethical use of AI in society (analogous to the field of biomedical ethics, which provides practical frameworks for ethical practice in medicine)” [32]. This definition emphasizes the novelty of the field, further highlighting the importance of biomedical ethics.

The emphasis on biomedical principles is consistent with current scientific and governmental efforts aimed at developing AI ethics frameworks and guidelines to ensure the ethical development, deployment, and use of AI technologies [29,33]. The biomedical principles mentioned in the definition of AI ethics refer to the well-known and established principles of medical ethics initially proposed by Beauchamp and Childress [34]. The 4 principles of autonomy, beneficence, nonmaleficence, and justice are considered fundamental to medical ethics, while most guidelines and frameworks on AI ethics do not specifically focus on ethical considerations regarding the development, implementation, or use of AI in medicine; the emphasis on these principles further reinforces their importance [30,35].

Although existing guidelines and frameworks aim to address various ethical concerns related to AI, such as privacy, bias, accountability, and transparency, it should be noted that they fail to provide a clear definition of AI ethics [30]. Given the rapid pace of advancements in AI technology and its increasing impact on society, the need for clear and consistent definitions of AI and AI ethics is becoming increasingly urgent [30]. To specifically address ethical considerations related to AI in medicine and medical practice, a definition of “medical AI ethics” has been proposed, which “is an interdisciplinary subfield of AI ethics concerned with the application of ethical principles and standards to the research, development, implementation, and use of AI technologies within the practice of medicine” [10]. This definition emphasizes the importance of principles regarding the use of AI in medicine, which is fundamental to this study.

**AI Ethics in Medical Education**

Although the need for teaching AI ethics in medical education is emphasized in scholarly literature, there is a lack of specification on relevant teaching content for AI ethics. In a recent scoping review, only a limited number of publications specifically focusing on the teaching of AI ethics as part of medical education were identified [10]. Although other publications acknowledge the importance of ethics in AI education, they do not provide specific content or guidance [36-39].

In one of the 2 identified publications specifically addressing AI ethics teaching content for medical education, 6 potential topics were defined: informed consent, bias, safety, transparency, patient privacy, and allocation [9]. The 6 teaching subjects were proposed to address the potential challenges related to the application of AI in medicine. For example, the anticipated challenge of informed consent highlights the importance of patient autonomy, potentially impeded by the lack of transparency or explainability in the decision-making of AI-based applications. Besides these 6 potential teaching subjects, the importance of teaching fairness and responsibility is emphasized by another publication that focuses on AI ethics education [11]. Furthermore, the importance of empathy has been emphasized in relation to the use of AI in medicine and the associated need to teach AI ethics [6].

A recurrent theme related to the teaching of AI ethics as part of medical education focuses on the principles of medical ethics according to Beauchamp and Childress (autonomy, beneficence, nonmaleficence, and justice) [10]. This emphasis is also echoed by existing guidelines and frameworks regarding AI ethics [30]. Additional recommendations on AI ethics teaching content include “explainability,” “liability,” and “accountability,” which are also considered important by available guidelines [30,40]. On the basis of the analysis of existing publications on teaching AI ethics in medical education, 12 potential subjects were considered for teaching AI ethics. The 12 identified potential teaching subjects for AI ethics in medical education are listed in Textbox 1.
Informed consent
Informed consent in the context of AI in medicine requires that patients be fully informed about treatment options and risks, necessitating a comprehensive understanding and explanation of AI technologies by physicians.

Bias
The use of AI in medicine may exhibit biases stemming from nonrepresentative data or structural conditions, leading to potential discrimination based on sex, age, or socioeconomic status.

Safety
The use of AI in medicine can have potentially harmful consequences for patients, necessitating a critical examination of the accuracy of AI-based applications and clear communication of their limitations.

Transparency
Transparency in AI-based medical applications is essential for understanding decision-making processes, influencing the quality and ethics of patient care, and maintaining trust, particularly in critical scenarios.

Privacy
Privacy not only refers to implementing technical data protection measures but also comprehensively understanding the ethical implications of handling sensitive patient data.

Allocation
In the context of AI in medicine, allocation refers to equitable access to technology and the impact of AI on equitable access to care.

Fairness
Fairness in AI ethics within medicine refers to ensuring equitable treatment for all patients regardless of their background. This encompasses the need for AI systems to be free from biases that may affect diagnosis, treatment recommendations, or patient outcomes.

Responsibility
Responsibility in the context of AI ethics in medicine emphasizes the importance of health care professionals and AI developers to using AI tools responsibly. This includes ensuring that these tools are safe, reliable, and used in a manner that benefits the patients.

Empathy
Empathy in the context of AI underscores the importance of maintaining the human aspect of health care, especially as AI technologies become more prevalent.

Explainability
Explainability in AI in medicine is closely linked to transparency and is important for understanding the AI-based decision-making process, affecting physician-patient relationships, and shared decision-making.

Liability
Liability in medical AI ethics concerns the potential for treatment errors related to the use of AI in the medical context. Questions on liability extend from potential users to health care institutions and AI developers.

Accountability
Accountability in medical AI involves understanding the associated limitations and competent oversight by medical professionals. This includes critically assessing AI errors and biases and ensuring accurate, informed, and ethical applications within medical decision-making. In addition, this accountability extends to continuously monitoring AI performance and adapting to evolving ethical and clinical standards in medical practice.

Objective
On the basis of a discussion and reflection theoretical analysis of the recommended teaching subjects on AI ethics informed by existing literature (as specified in the AI Ethics in Medical Education section), this study aims to introduce a set of ethical principles for “medical AI ethics.” As the proposed AI ethics teaching subjects for medical education in the existing scientific literature primarily focus on the challenges associated with the use of AI in medicine, they fail to acknowledge the broader implications of foundational ethical principles. By concentrating on a principle-based approach to AI ethics, this paper aims to address the gap in the existing scientific literature, serving as a foundational framework for AI ethics teaching content in medical education.

Theoretical Analysis of Recommended AI Ethics Teaching Subjects in Medical Education

Overview
Ethics commonly relies on principles as foundational guidelines for decision-making and behavior. The 4 foundational principles of medical ethics—autonomy, beneficence, nonmaleficence, and justice—are highly relevant in the context of teaching ethics in medical education [41].

While these 4 principles have been an integral part of current scientific publications on AI ethics in medical education, the
recommended teaching subjects are mainly derived from the anticipated challenges associated with the use of AI in medicine [10]. Addressing these challenges is important for fostering a comprehensive understanding regarding the use of AI in medicine. However, this approach does not fully capture the multidisciplinary and interdisciplinary nature of this field. The complexity of AI ethics in medicine extends beyond these anticipated challenges, encompassing a wide range of disciplines such as law, medicine, ethics, and computer science. For example, the proposed teaching subject of “informed consent” warrants a detailed analysis to exemplify the high level of interdisciplinarity present in AI ethics, intersecting with each of the other proposed teaching subjects. This interconnection results in a substantial overlap, which can challenge the establishment of clear distinctions between the different areas of AI ethics.

The methodology of this study is anchored in a theoretical approach, building upon a previous comprehensive scoping review of the existing literature on teaching AI ethics in medical education [10]. This also includes the consideration of relevant guidelines and frameworks regarding the ethics of AI, resulting in the identification of 12 potential teaching subjects for AI ethics as detailed in Textbox 1. To exemplify the high level of interdisciplinarity present in AI ethics by focusing on the subject of “informed consent,” the publications included in the scoping review, including the proposed challenges associated with the use of AI in medicine, were re-evaluated. This theoretical analysis provides the foundation for the development of the principles of medical AI ethics presented in the Medical AI Ethics section. The theoretical basis of the proposed principle-based approach to AI ethics is further strengthened by our expertise as we specialize in the ethical use of AI in medical and public health contexts. This background informs the depth and rigor of the analysis, ensuring that the developed framework is both relevant and grounded in practical ethical considerations in these fields. The theoretical methodology we used is characterized by a focus on conceptual development and theoretical insights rather than empirical testing or data collection.

**Informed Consent**

**Overview**

Informed consent represents an important development in medical ethics and patient rights, representing a departure from the historically paternalistic nature of medical practice [42]. In earlier medical paradigms, decision-making was predominantly physician driven, with minimal patient involvement. This approach, often paternalistic, assumes the primacy of the physician’s judgment, potentially leading to interventions conducted without comprehensive patient understanding or consent [42].

The development and integration of informed consent into medical practice represents a substantial cultural and ethical transition toward acknowledging and upholding patient autonomy. Central to this evolution is the concept of shared decision-making (SDM), a collaborative process that involves physicians and patients jointly making treatment decisions. SDM encompasses a thorough discussion of available treatment options, including their benefits and risks, and considers patient values, preferences, and circumstances [42,43]. This method positions patients as active participants in their health care journey rather than as passive recipients of medical decisions.

In this context, informed consent is pivotal in facilitating SDM, as it ensures that patients are not only informed of their medical choices but also engaged in selecting options that resonate with their personal health goals and values. This approach transforms the traditional physician-patient relationship into a partnership, where decisions are mutually agreed upon, thereby honoring the patient’s right to self-determination. It also fosters a deeper level of trust and respect within the physician-patient relationship.

As a result, informed consent serves more than just a legal requirement to minimize liabilities; it is a crucial aspect of patient-centered care and a fundamental element of ethical medical practice. This signifies the transition from a paternalistic approach to one that emphasizes patient autonomy and upholds the principles of SDM.

**Informed Consent in the Context of AI in Medicine**

Regarding the development, implementation, and use of AI in medicine, the concept of informed consent warrants a comprehensive introduction owing to the technical complexities inherent to AI. AI systems, particularly those used in diagnostics and treatment recommendations such as ML, often involve algorithms that might be nontransparent to both patients and health care professionals. This lack of transparency presents a substantial challenge to the conventional process of informed consent, complicating the task of understanding and communicating how an AI-based application formulates recommendations [44].

Moreover, the development of AI-based applications involves extensive data sets, raising concerns regarding data privacy and the potential for expropriation of personal health data [9]. These issues necessitate clear communication with patients throughout the physician-patient relationship and during the process of ensuring informed consent. It is imperative that patients are adequately informed about not only the advantages and risks associated with AI-assisted treatments but also the manner in which their data are used, protected, and stored [45]. With the increasing integration of AI in medicine and health care, the process of obtaining informed consent must be adapted to meet these challenges, thereby ensuring that patients retain control over their health care decisions in an environment increasingly influenced by AI.

**Intersections of Informed Consent With Key AI Ethics Teaching Subjects**

**Overview**

This section aims to underscore interdisciplinarity and intersectionality among the recommended teaching subjects in AI ethics, as outlined in the Informed Consent section, with informed consent serving as a representative example. Focusing on these intersections, this section highlights the importance of an integrated educational approach in the context of medical AI ethics. Such an approach acknowledges that topics such as
bias, privacy, and transparency, among others, are not merely isolated subjects but instead require a comprehensive, holistic evaluation. Embracing this integrated perspective is important for a comprehensive understanding of AI ethics in medical practice and education, underscoring the need to re-evaluate and potentially refine current teaching recommendations. To effectively illustrate the interdisciplinarity and interconnectedness of frequently recommended teaching subjects for AI ethics in medical education, “informed consent” should be discussed in the context of 5 frequently proposed teaching subjects: bias, safety, transparency, privacy, and liability.

**Bias**

To enable patients to make informed decisions when AI-based applications are used in their treatment, it is important to address the possibility of bias inherent in these technologies. Informed consent in this context requires the awareness and understanding of potential biases in AI decision-making processes [46]. For instance, a diagnostic AI-based application might exhibit varying levels of accuracy across different demographic groups, potentially owing to data representation issues [21]. Patients must be informed of such disparities in accuracy as this information is vital for them to consent to the use of AI in their treatment.

**Safety**

The safety of AI-based applications in medicine is a critical component of informed consent for medical treatment recommendations involving AI. Patients must be clearly informed about the potential risks associated with AI-driven medical decisions, including the possibility of erroneous outcomes such as false positives or negatives [47]. This comprehensive understanding of the safety profile of AI-assisted treatments is essential for patients to make informed decisions about their care. Being informed and knowledgeable about the limitations and risks of AI technologies ensures that patients can weigh these factors against potential benefits when consenting to AI use in their treatment.

**Transparency**

Transparency in AI systems is important not only for patients but also for physicians, who serve as the primary receivers and communicators of AI-driven medical information. A clear understanding of how AI-based applications work, particularly how decision-making processes are performed, is required for physicians to effectively communicate with their patients [48]. Such informed communication is a fundamental aspect of the informed consent process, fostering a deeper understanding and trust within the physician-patient relationship [49]. When patients receive comprehensive and transparent information from their trusted health care providers, they enhance their engagement and participation in decision-making. Therefore, transparency in AI goes beyond technical clarity and is crucial for fostering a strong physician-patient relationship, ensuring that informed consent is based on a shared understanding of the potential risks and benefits associated with AI-assisted treatments [50].

**Privacy**

The process of obtaining informed consent for AI-based medical treatment recommendations should include data privacy. It is important for patients to be informed about the use, access, and protection of their data. Ensuring that patients understand how their personal health data are used, who has access to it, and the measures in place to protect it is a key component of the informed consent process [51]. This comprehensive disclosure and transparency regarding data handling are vital for maintaining the integrity of the physician-patient relationship and for upholding the ethical standards of medical practice in the era of AI.

**Liability**

Regarding the use of AI in medicine, it is imperative to address the concept of liability in the informed consent process. Patients should be clearly informed of the potential for errors and liability issues associated with AI-driven medical decisions [52]. This conversation should entail a discussion on who bears responsibility, including the liability of physicians, if an AI system malfunctions or leads to incorrect medical outcomes such as misdiagnoses or inappropriate treatment plans. The explicit clarification of liability, particularly the role and responsibility of health care providers in conjunction with AI, is important for helping patients understand the potential risks involved [53]. This understanding is a key component of a comprehensive informed consent process that directly affects the patients’ trust in AI and their treating physicians. By transparently addressing these liability concerns, including the physicians’ responsibilities, health care providers can reinforce the integrity of the physician-patient relationship and uphold the ethical standards of medical practice in an AI-integrated health care environment [53].

**Medical AI Ethics**

**Overview**

The high degree of interdisciplinarity and intersectionality in AI ethics, as detailed in the previous section, highlights potential conflicts in teaching AI ethics based solely on the anticipated challenges associated with the implementation and use of AI in medicine. This complexity underscores the necessity of adopting a principle-based approach to AI ethics education, mirroring established pedagogical frameworks in medical ethics education [41].

In the context of traditional medical ethics education, the emphasis on foundational principles provides a broad and adaptable framework that is essential for understanding and addressing complex ethical dilemmas. This approach facilitates the holistic comprehension of ethical issues, offering the flexibility to accommodate the diverse and evolving nature of medical scenarios. Similarly, when considering AI in medicine, a focus on core ethical principles rather than solely on specific challenges lays the groundwork for a robust and comprehensive educational strategy. Future medical professionals should be equipped with a deeper and more nuanced understanding of ethical decision-making by emphasizing ethical principles in the context of implementing and using AI in medicine. This principle-based approach ensures that medical ethics education
remains relevant and responsive to the dynamic landscape of AI integration in medicine. The goal is for medical students to be able to effectively navigate the ethical complexities associated with AI technologies in medicine, not just focusing on potential challenges but also emphasizing the ethical values that are essential to medical practice.

Owing to the paramount importance and relevance of the 4 principles of medical ethics formulated by Beauchamp and Childress [34], the principles of autonomy, beneficence, nonmaleficence, and justice should provide the essential foundation for medical AI ethics. These 4 principles are subsequently introduced based on existing scholarly discourse, focusing on the use of AI in medicine, with an emphasis on medical education.

Traditional medical practices have predominantly focused on individual relationships between physicians and patients. However, modern health care increasingly necessitates considering broader aspects such as cost-effectiveness, resource allocation, and proportionality, especially in light of financial constraints. A prominent illustration of these evolving dynamics in medical practice is the COVID-19 pandemic. This global health crisis underscored the critical importance of public health considerations and highlighted extensive interdisciplinarity and interconnectivity within the field of medicine. The COVID-19 pandemic has highlighted the importance of balancing individual patient care with broader public health measures [54]. It demonstrated how medical decisions are not made in isolation but are profoundly influenced by factors such as resource availability, health care infrastructure, and broader societal implications. This scenario emphasizes the crucial role of public health principles in informing medical practices, particularly in crises. The pandemic also illustrates the necessity of integrating insights from various disciplines, including epidemiology, health economics, and ethics into medical decision-making.

Given the anticipated impact of AI on the field of medicine, which extends beyond the traditional concept of medical practice owing to its inherent interdisciplinarity and complexity, ethical considerations must be adapted accordingly. The scope of AI in medicine introduces novel ethical dimensions that require a broader framework for ethical analysis. Therefore, the integration of 3 principles of public health ethics—efficiency, common good orientation, and proportionality—is proposed along with the established principles of medical ethics to form a comprehensive foundation for medical AI ethics [55-58].

Similar to the principles of medical ethics outlined by Beauchamp and Childress [34], each principle of public health ethics is examined in subsequent sections with a specific focus on its relevance to AI in medical practice and education. While the principles of public health ethics may not be as established or universally agreed upon as those of medical ethics, their inclusion provides a suitable framework to address the unique challenges posed by AI in medicine and health care. This extended ethical framework aims to provide a more comprehensive understanding of the role and implications of AI in medicine, ensuring that future medical professionals are equipped to make ethically sound decisions in increasingly AI-integrated medical practice. The proposed principles of AI ethics for medical education are presented in Figure 1.

**Figure 1.** The principles of medical artificial intelligence (AI) ethics for medical education.

**Autonomy**

The principle of autonomy in medical ethics emphasizes the right to make independent decisions regarding health care [34]. This principle recognizes an individual’s capacity for self-determination and personal choice, affirming that patients have the authority to provide or withhold consent for medical treatment. Respecting autonomy in medical practice involves providing patients with sufficient information, ensuring comprehension, and facilitating independent decision-making [59]. This respect for autonomy is closely tied to the principle of informed consent, which ensures that patients actively participate in decisions regarding their care and treatment.
In the context of using AI in medicine, particularly in diagnostics and treatment recommendations, technology introduces new challenges and opportunities to maintain patient autonomy [60]. For example, when using AI-based diagnostic applications, it is crucial to inform patients about how these tools impact their health care decisions, ensuring that informed consent is comprehensive. Equally important is equipping physicians with the knowledge to balance AI-generated insights with their clinical expertise, thus upholding both patient and physician autonomy in decision-making processes. The incorporation of AI into health care decision-making can affect the presentation and comprehension of options by patients. Ensuring that patients retain their autonomous decision-making power in an AI-driven environment requires the careful consideration of how the information is communicated and understood [60]. Autonomy in this context extends to ensuring that patients have a clear understanding of AI interventions and their capabilities, limitations, and impact on personal health decisions. Moreover, the principle of autonomy extends to physicians. If AI increasingly assists in medical decision-making, it is imperative that physicians remain empowered to make independent professional judgments, balancing AI insights with their clinical expertise and ethical considerations.

The principle of autonomy addresses several anticipated challenges and recommends teaching subjects on AI ethics in medical education. For example, in the context of informed consent, autonomy ensures that patients are fully aware of the role and limitations of AI in their treatment, including potential bias and safety concerns. Autonomy also involves clear communication regarding data privacy, ensuring that patients understand how their data are used in AI systems. In the context of using AI in medicine, autonomy is not limited to the patient’s understanding and decision-making; it also encompasses the physician’s ability to make independent judgments informed by, but not solely reliant on, AI-driven data. This dual focus preserves the integrity of clinical decision-making and respects both the patient’s and the physician’s autonomous roles. Furthermore, transparency and explainability in AI systems are fundamental to ensure that patients autonomously understand and evaluate AI-driven health care choices. Autonomy acts as a guiding principle that addresses these challenges, ensuring that patient rights and self-governance remain central to the increasingly AI-integrated landscape of medical practice. This principle also extends to the equitable allocation of medical resources and fairness in treatment decisions, where an autonomous choice must be informed by unbiased AI recommendations. This comprehensive approach to autonomy in AI ethics education underscores the need for a balanced consideration of both patient and physician perspectives to ensure ethical integrity in the application of AI in medicine.

**Beneficence**

The principle of beneficence, a fundamental aspect of medical ethics, underscores the responsibility of health care providers, including physicians, to act in the best interests of patients [61]. This principle is the basis of the ethical framework guiding health care delivery and promoting actions that enhance patient well-being and welfare [34]. In medical practice, beneficence guides physicians to consider the actual benefits of medical interventions, extending from the sole minimization of potential harm. Therefore, this principle encompasses a broader responsibility toward enhancing the overall quality of life of the patient, affirming that every medical decision should contribute positively to the holistic well-being of the patient [50].

The principle of beneficence is paramount in the application of AI in medicine, such as through predictive analytics and personalized medicine. Although promising, AI-based applications must be critically evaluated for their efficacy and safety to ensure alignment with the overarching goal of promoting patient well-being, which reflects the true essence of beneficence in medical practice [62]. In addition, it is crucial to ensure that AI-based applications align with the broader goals of patient care, emphasizing not only clinical outcomes but also patient quality of life and overall well-being. Such an approach should consider individual social backgrounds and personal circumstances, ensuring that AI-driven health care focuses on the diverse needs of each patient [50].

In the context of AI ethics and medical education, beneficence emphasizes the importance of developing, implementing, and using AI applications designed with the primary aim of improving patient outcomes. This includes addressing potential biases in AI algorithms that could negatively impact patient care, ensure patient safety, and maintain transparency in the AI decision-making processes. Therefore, the principle of beneficence guides the ethical application of AI in medicine, ensuring that these advancements aim to maximize patient benefits and well-being, consistent with the overarching goals of medical practice.

**Nonmaleficence**

Although the principle of nonmaleficence also focuses on ensuring the best possible treatment for patients and aligning all actions accordingly, it emphasizes that health care professionals should do no harm [34]. This principle is complementary to the principle of beneficence, and it aims not only to prevent harm but also to proactively avoid and reduce risks associated with medical care. Nonmaleficence requires that the risks of any medical intervention are carefully weighed against their potential benefits and actions that could cause harm are avoided. This principle underlines the responsibility of health care providers to ensure that any treatment or medical advice does not adversely affect a patient’s health.

The potential risks of using AI in medicine, such as misdiagnosis, algorithmic biases, and data security breaches, reinforce the relevance of the principle of nonmaleficence. To ensure nonmaleficence, the rigorous testing and validation of AI systems, ongoing monitoring for adverse outcomes, and commitment to addressing any safety concerns are crucial [62]. Moreover, this commitment extends to the ethical development and deployment of AI technology. It involves actively working to mitigate risks, such as biases in training data, that could lead to unequal or unfair treatment outcomes [50].

To raise the awareness of potential conflicts with the principle of nonmaleficence regarding the use of AI in medicine, medical education should focus on the ethical design, development, and
deployment of AI applications in medicine. Therefore, nonmaleficence is an important part of medical AI ethics, emphasizing the need to ensure the accuracy and reliability of treatment recommendations originating from the use of AI-based applications in medicine. Teaching content on nonmaleficence addresses various anticipated challenges regarding the use of AI in medicine, such as safety, privacy, bias, and transparency. By adhering to the overarching principle of nonmaleficence, physicians can navigate the ethical challenges posed by AI in medicine, ensuring that the technology is used in ways that prioritize patient safety and harm reduction.

**Justice**

The principle of justice in medical ethics, as outlined by Beauchamp and Childress [34], is concerned with ensuring fair and equal treatment for all patients regardless of their socioeconomic status, background, or circumstances. This principle emphasizes the importance of fairness in the distribution of resources and access to health care services. In practical medical settings, justice can be translated into unbiased decision-making, equal opportunity for treatment, and eradication of any form of discrimination.

Justice is an important aspect regarding the use of AI in medicine. Owing to the risk of bias due to unrepresentative training data, for example, treatment recommendations from the use of AI in medicine could lead to disadvantages for different groups or individuals, directly conflicting with the principle of justice [50]. Furthermore, access to the technology itself could be limited, for example, by economic means, thereby potentially perpetuating existing inequalities in access to advanced medical technologies [35]. This potential for injustice can be further exacerbated if an increasing prevalence of AI in medical practice is anticipated.

Owing to the substantial risk of injustice with the use of AI in medicine, medical education should include teaching the principle of justice in the context of AI. Focusing on the equitable availability and use of AI technologies, future physicians should be trained to recognize and address the potential inequities that AI might introduce or perpetuate. Therefore, teaching the principle of justice, extending from traditional medical ethics education, can serve as a foundation to address anticipated challenges such as allocation, bias, fairness, liability, and accountability. For instance, when considering liability and accountability, justice refers to ensuring that patients are not disproportionately affected by errors or failures in AI systems. It involves advocating for systems that hold developers and health care providers responsible for potential technological malfunctions, ensuring that accountability measures are in place to protect all patients from potential harm or injustice, especially those in vulnerable or marginalized groups [53].

**Efficiency**

Efficiency within public health ethics underscores the strategic use of resources to maximize health benefits for the population [57]. This principle is not only solely an economic concern but also a moral imperative to ensure the equitable and judicious use of medical technologies and services. Ethical considerations regarding the principle of efficiency are especially relevant in health care settings where resources are limited and demand is high, as exemplified in the context of the COVID-19 pandemic. Owing to the capabilities of AI in medicine with the potential to enhance the efficiency of medical services through faster and more accurate diagnostics, it is crucial to consider the ethical implications of these developments [19]. The ability of AI to rapidly analyze large data sets can greatly enhance the speed of diagnostic procedures, which could result in more timely patient care and improved treatment choices that are more precise. However, this benefit is contingent on the data quality. Poor-quality data can result in AI models that incorrectly predict outcomes based on artifacts in the data rather than actual clinical results [21]. Therefore, the ethical use of AI in health care must include rigorous validation of the data quality to ensure accurate and reliable outcomes. For example, physicians must balance the efficiency gains offered by AI with the need for clinical judgment and personalized patient care and upholding and maintaining the quality of physician-patient relationships [63].

Teaching the principle of efficiency in the context of AI ethics education should focus on the balance between technology-driven efficiency and patient-centered care. Future physicians need to understand how to leverage AI to optimize health care delivery without compromising quality of care. Therefore, teaching the principle of efficiency highlights the anticipated challenges related to a lack of empathy. It is imperative to ensure that the pursuit of efficiency through AI does not lead to the depersonalization of patient care. Empathy remains a crucial aspect of health care, and AI systems should be used to enhance, rather than replace, the human elements of patient interaction and care.

**Common Good Orientation**

Common good orientation is a guiding principle of public health ethics, aiming to improve the collective well-being and health of the community or population as a whole [58]. This principle extends the focus of individual patients, emphasizing the interconnectedness between individual and public health. This involves considering the wider impacts of health care interventions and prioritizing actions that promote the health and welfare of the public.

The principle of common good orientation in the context of AI, crucial in guiding the integration of technology into medical practice, calls for a delicate balance between individual patient benefits and the collective well-being of the community. It is essential to recognize how AI in medicine can address or potentially exacerbate health disparities [64]. The ability of AI to process and analyze data can be harnessed to identify and address gaps in health care delivery, offering insights into underserved populations and tailoring interventions to meet their specific needs. Conversely, if not carefully managed, AI could unintentionally increase these disparities by favoring populations with better access to the technology. This duality underscores the need for AI advancements in health care to contribute positively and equitably to public health, promoting fairness in health care access and outcomes. It is important to note that the selective application of AI not only undermines the principle of common good orientation but also risks creating...
a perception of elitism in the medical profession. Such a scenario could harm the reputation of the medical field, rendering it as unequally benefiting certain populations. Furthermore, using AI in medical practice could potentially lead to events where patients are harmed, for example, through biased decision-making or errors made by users. This could potentially lead to a negative perception of AI within the broader population, which in turn may result in a general unwillingness or resistance to adopting AI technologies. This hesitance could directly conflict with the principle of common good orientation, as it hinders the widespread and equitable implementation of AI that could benefit the entire community [25].

Teaching the principle of common good orientation in the context of AI ethics in medical education underscores the importance of developing, implementing, and using AI technologies in ways that serve a wider community not just the individual patient. This includes understanding the potential of AI in managing public health crises such as pandemics. Medical education based on the principle of common good orientation emphasizes aspects of safety, transparency, allocation, and responsibility, which are important to best prepare for potential challenges through AI in medicine and associated ethical considerations.

Proportionality

The principle of proportionality in public health ethics necessitates a balanced approach to medical interventions that weighs benefits against risks [57]. Therefore, this principle can be applied to ensure that the measures taken, such as medical interventions, are proportional to the health risks that they aim to mitigate. In medicine, proportionality is important in decision-making, ensuring that the intervention aligns with the expected health outcomes.

In medical practice, the principle of proportionality is important when considering the integration of AI technology to balance benefits against potential risks for individual patients and the broader population. This principle necessitates a careful assessment of the role of AI, particularly in ensuring equitable resource distribution and maintaining public trust [25]. For instance, when using AI for diagnostics, it is necessary to evaluate the accuracy and effectiveness of the technology against risks, such as misdiagnosis or overreliance on AI. This evaluation should consider not only the immediate impact on individual patients but also the broader implications for health care resources and community trust. In the critical area of resource allocation within health care, the use of AI holds substantial promise in enhancing the efficiency and effectiveness of distributing limited medical resources [63]. However, it is essential to guard against the risk of AI systems inadvertently perpetuating existing biases or failing to address the diverse needs of different patient groups. This calls for a transparent, community-engaged approach to the development and deployment of AI in health care, ensuring that AI recommendations do not unfairly disadvantage any patient group [24]. By adhering to the principle of proportionality, health care providers can better navigate the ethical complexities of using AI, ensuring that its application is not only technologically sound but also ethically responsible, both at the individual patient level and in the wider context of public health.

The principle of proportionality can be helpful for future physicians to comprehend the anticipated challenges of AI in medicine, particularly regarding the aspects of allocation. This principle also addresses other anticipated challenges such as transparency and explainability to understand how decisions are made and whether the overall population is considered, ensuring that recommendations are reasonable.

Discussion

Overview

The integration of AI in medicine necessitates a nuanced approach to ethics education that addresses the unique challenges and opportunities introduced by this technology. By exploring public health and medical ethics principles, medical AI ethics offers a comprehensive framework for guiding future physicians in this complex landscape. The proposed teaching of medical AI ethics in medical education emphasizes the importance of ethical principles rather than focusing solely on anticipated challenges, aiming to foster a deeper understanding of potential ethical considerations and enable adaptation in the light of rapid technological advancements.

Given the dynamic nature of AI and the associated rapid technological advancements, for example, as demonstrated by AI-based chat applications such as ChatGPT, ethical considerations need to be continually adapted [65]. The need for timely adaptation challenges traditional ethics education in medicine, which may not account for the current use of AI in medicine. Traditional ethics education primarily focuses on the 4 principles of medical ethics as formulated by Beauchamp and Childress [41]. While these principles can provide valuable guidance in the age of AI in medicine and are therefore foundational to the proposed medical AI ethics education, adaptation is needed to reflect the complexities and challenges introduced by the implementation and use of AI in medicine and medical practice.

The high level of intersectionality and interdisciplinarity inherited by the implementation and use of AI in medicine highlights the importance of a principle-based approach rather than solely focusing on anticipated challenges. While the proposed ethical principles also show a high level of interconnectivity, the chosen educational approach aims to encourage a more nuanced understanding, not limited to specific anticipated challenges but rather to enable future physicians to adapt to the changing landscape associated with the use of AI in medicine, facilitating the consideration of multiple ethical dimensions simultaneously. In addition to the proposed principles, medical education should incorporate practical case studies and simulations to reflect real-world scenarios. For example, applying AI to patient triage during health emergencies such as the COVID-19 pandemic can offer practical contexts for students. This approach would not only enhance their understanding of ethical principles but also prepare them for decision-making in complex, real-life medical situations influenced by AI. It is important for future physicians to
understand the balance between the potential benefits of AI and the ethical implications of its use, particularly in scenarios in which biased algorithms could lead to unequal treatment of diverse patient groups. Therefore, a comprehensive curriculum that includes both theoretical knowledge and practical applications is essential to cultivate ethically informed medical professionals.

An in-depth and interdisciplinary understanding of ethics is important in the dynamic field of medical AI. This importance is underscored by the fact that the integration of AI into medical education may not always keep pace with rapid advancements in medical practice. A focus on ethical principles rather than solely on specific challenges of AI use in medicine aims to prepare medical students for various scenarios in the medical context. This approach maintains relevance even if the AI applications used in education are not representative of the latest state-of-the-art developments in medical AI. The principle-based approach to AI ethics offers broader applicability and reduces dependence on the most recent AI technologies, potentially benefiting medical schools with limited financial resources. In addition, AI products for teaching, often sourced from third parties and guided by cost considerations, may pose unique challenges such as the risk of bias or rapid obsolescence [66,67].

This necessitates awareness, among medical students, of the potential ethical issues associated with these tools. By emphasizing a principle-based approach to AI ethics, educators can equip students with the necessary understanding to navigate the evolving landscape of AI in medicine, fostering adaptability and ethical sensitivity in future medical professionals. This adaptability is crucial to ensure that future physicians are prepared for the ethical dilemmas they may encounter in a rapidly evolving AI landscape.

In the applicability of the principle-based approach to AI ethics, the paramount importance of AI-based chat applications such as ChatGPT must be assumed [68]. As ChatGPT demonstrated extensive medical knowledge, as exemplified by its ability to pass the written part of the United States Medical Licensing Exam, AI-based chat applications offer new opportunities for medical education and medical students, such as in simulated patient interactions and case study analysis [69,70]. However, as ChatGPT was not explicitly developed for use in the medical context and, for example, does not adhere to stringent medical device regulations, it raises new ethical challenges. This becomes particularly evident, as AI-based chat applications can hallucinate and might not provide correct medical information due to improper “prompting” [70]. The limitations of ChatGPT, such as inaccurate or misleading medical information, necessitate an awareness of not only the technical limitations but also the associated ethical considerations. This reinforces the importance of a principle-based approach to AI ethics in medical education, emphasizing the importance of critically reflecting on and evaluating any use of AI in medicine. Awareness of potential ethical considerations regarding AI-based chat applications also extends from the provision of medical knowledge to a broader medical context, such as scientific research [71]. For example, if AI-based chat applications such as ChatGPT are used for medical research, medical education should facilitate an understanding of how this could impact research integrity or potentially interfere with the existing ethical standards [71]. Medical education should prepare students to navigate through these complexities, ensuring the ethical integration of AI in practice and research.

Although the integration of public health ethics principles as part of medical AI ethics offers a comprehensive approach for teaching AI ethics in the medical setting, it is important to recognize that the field of public health ethics is still evolving [72]. Unlike the well-established principles of medical ethics proposed by Beauchamp and Childress [34], public health ethics principles such as efficiency, common good orientation, and proportionality are not universally agreed upon or applied consistently across different contexts. This lack of standardization presents a challenge for formulating a universally applicable ethical framework for AI in medicine. Furthermore, the interdisciplinary nature of public health ethics, encompassing the aspects of sociology, economics, and political science, adds to the complexity of integrating these principles into medical AI ethics education. This complexity requires careful consideration during curriculum development to ensure that these principles are taught in a manner that is both relevant and applicable to medical students. Moreover, the rapidly changing landscape of AI technology necessitates a dynamic approach to ethics education in which principles and guidelines are continuously revisited and updated. This need for adaptability may challenge the traditional formats of medical education, calling for innovative pedagogical approaches to ensure that future physicians are adequately prepared for the ethical complexities of AI-integrated medical practice.

**Limitations**

This study and the proposed theoretical foundation to medical AI ethics is subject to several limitations that need to be considered. Continuous evolution in the field of AI presents substantial challenges for the development of static ethical guidelines and frameworks for medical education. The dynamic nature of AI technology underscores the need for an adaptable and responsive ethical framework in medical education, particularly in the context of public health ethics, where principles are still developing and gaining consensus. Given that new advancements, for example, as exemplified by AI-based chat applications such as ChatGPT, cannot be foreseen and that the capabilities of AI and AI-based applications in medicine are anticipated to expand, continuous updates of existing educational frameworks and content are required.

Furthermore, the applicability and relevance of ethical principles as a part of medical AI ethics education may vary across cultural and health care settings. Different regions may have varying access to AI technologies, and cultural values may influence the perceptions of integrating and using AI in the medical setting. This variability could impact the universality of the proposed ethical framework and limit the applicability of teaching medical AI ethics as a part of medical education.

Moreover, integrating new teaching content into medical curricula is challenging due to the need for time-intensive accreditation processes and extensive teaching content. The integration of new teaching content such as medical AI ethics education requires careful planning to ensure that future
physicians are adequately prepared and not overwhelmed by information. In addition, limited access to instructors knowledgeable in ethics, medicine, and AI may pose a challenge to implementing the proposed teaching of medical AI ethics, as these experts may not be available in most institutions.

Conclusions
This study highlights the imperative need for medical AI ethics education and the integration of a comprehensive set of ethical principles into medical education to prepare physicians for the ethical challenges posed by AI in medicine. As the advancement of AI technologies in medicine is expected to increase, it is essential for medical ethics education to adapt and evolve accordingly to keep pace with these developments. Educational institutions should take proactive steps to update their curricula, ensuring that future medical professionals are not only aware of the ethical dimensions of AI in medicine but also equipped to make informed ethical decisions in their practice. The principles discussed, drawn from both traditional medical and public health ethics, provide a multidimensional framework for understanding and navigating the ethical landscape associated with the use of AI in medicine.

Given the rapid advancements in the field of AI, it is essential that these ethical guidelines be regularly revisited and updated to remain relevant in the context of medical education. The proposed dynamic approach, with an emphasis on ethical principles, aims to ensure that medical professionals not only are equipped to use AI in ways that enhance patient care but also uphold the highest ethical standards. Future research is needed to develop problem-based and competency-oriented learning objectives and educational content for medical AI ethics and implementation and validation.

Conflicts of Interest
None declared.

References
70. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]


Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
ML: machine learning
NLP: natural language processing
SDM: shared decision-making

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 11.12.23; peer-reviewed by KH Miller, Z Khlaif; comments to author 29.12.23; revised version received 02.01.24; accepted 29.01.24; published 09.02.24.

Please cite as:
Weidener L, Fischer M
Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education
JMIR Med Educ 2024;10:e55368
URL: https://mededu.jmir.org/2024/1/e55368
doi:10.2196/55368
PMID:38285931

©Lukas Weidener, Michael Fischer. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Sharing Digital Health Educational Resources in a One-Stop Shop Portal: Tutorial on the Catalog and Index of Digital Health Teaching Resources (CIDHR) Semantic Search Engine

Julien Grosjean1,2*, PhD; Arriel Benis3,4*, PhD; Jean-Charles Dufour5, MD, PhD; Émeline Lejeune1, BSc; Flavien Disson1, MSc; Badisse Dahamma1,2, MSc; Hélène Cieslik1, MSc; Romain Léguillon1,2,6, MSc, PharmD; Matthieu Faure7, PhD; Frank Dufour8, PhD; Pascal Staccini8, MD, PhD; Stéfan Jacques Darmoni1,2,4, MD, PhD

1Department of Digital Health, Rouen University Hospital, Rouen, France
2LIMICS, INSERM U1142, Sorbonne Université, Paris, France
3Department of Digital Medical Technologies, Holon Institute of Technology, Holon, Israel
4European Federation for Medical Informatics, Le Mont-sur-Lausanne, Switzerland
5SESSTIM, Aix Marseille Univ, APHM, INSERM, IRD, Hop Timone, BioSTIC, Marseille, France
6Department of Pharmacy, Rouen University Hospital, Rouen, France
7Délégation du Numérique en Santé, Paris, France
8RETINES, Université de Nice Côte d’Azur, Nice, France
*these authors contributed equally

Corresponding Author:
Julien Grosjean, PhD
Department of Digital Health
Rouen University Hospital
1, rue de Germont
Rouen, 76031
France
Phone: 33 232885616
Email: julien.grosjean@chu-rouen.fr

Abstract

Background: Access to reliable and accurate digital health web-based resources is crucial. However, the lack of dedicated search engines for non-English languages, such as French, is a significant obstacle in this field. Thus, we developed and implemented a multilingual, multiterminology semantic search engine called Catalog and Index of Digital Health Teaching Resources (CIDHR). CIDHR is freely accessible to everyone, with a focus on French-speaking resources. CIDHR has been initiated to provide validated, high-quality content tailored to the specific needs of each user profile, be it students or professionals.

Objective: This study’s primary aim in developing and implementing the CIDHR is to improve knowledge sharing and spreading in digital health and health informatics and expand the health-related educational community, primarily French speaking but also in other languages. We intend to support the continuous development of initial (ie, bachelor level), advanced (ie, master and doctoral levels), and continuing training (ie, professionals and postgraduate levels) in digital health for health and social work fields. The main objective is to describe the development and implementation of CIDHR. The hypothesis guiding this research is that controlled vocabularies dedicated to medical informatics and digital health, such as the Medical Informatics Multilingual Ontology (MIMO) and the concepts structuring the French National Referential on Digital Health (FNRDH), to index digital health teaching and learning resources, are effectively increasing the availability and accessibility of these resources to medical students and other health care professionals.

Methods: First, resource identification is processed by medical librarians from websites and scientific sources preselected and validated by domain experts and surveyed every week. Then, based on MIMO and FNRDH, the educational resources are indexed for each related knowledge domain. The same resources are also tagged with relevant academic and professional experience levels. Afterward, the indexed resources are shared with the digital health teaching and learning community. The last step consists of assessing CIDHR by obtaining informal feedback from users.

https://mededu.jmir.org/2024/1/e48393

JMIR Med Educ 2024 | vol. 10 | e48393 | p.47
(page number not for citation purposes)
Results: Resource identification and evaluation processes were executed by a dedicated team of medical librarians, aiming to collect and curate an extensive collection of digital health teaching and learning resources. The resources that successfully passed the evaluation process were promptly included in CIDHR. These resources were diligently indexed (with MIMO and FNRDH) and tagged for the study field and degree level. By October 2023, a total of 371 indexed resources were available on a dedicated portal.

Conclusions: CIDHR is a multilingual digital health education semantic search engine and platform that aims to increase the accessibility of educational resources to the broader health care–related community. It focuses on making resources “findable,” “accessible,” “interoperable,” and “reusable” by using a one-stop shop portal approach. CIDHR has and will have an essential role in increasing digital health literacy.

(JMIR Med Educ 2024;10:e48393) doi:10.2196/48393

KEYWORDS
digital health; medical informatics; medical education; search engine; knowledge management; semantic web; language; teaching; vocabulary; controlled; students; educational personnel; French; curriculum

Introduction

Background

Medicine, health care, and wellness will become increasingly digitized. Thus, digital technologies are more than ever taking a pivotal position in clinical practice, making it crucial to educate future professionals to efficiently grasp digital health and health informatics [1,2]. The World Health Organization views digital health as “a broad umbrella term encompassing eHealth, mHealth, as well as emerging areas, such as the use of advanced computing sciences in big data, genomics, and artificial intelligence.” The World Health Organization affirmed that to strengthen health systems using digital health technologies, finding ways to build capacity and creating a digitally capable health workforce should be key objectives [3,4].

The integration of digital technologies has brought about significant changes in the realm of health professions education. Our research identified various digital education–related inquiries, culminating in a comprehensive and diverse research agenda. We proposed a conceptual framework to assist educators and researchers in developing, designing, and studying digital education. However, we acknowledge the need for further data from lower- and middle-income countries [5].

In 2022, the Delegation of Digital Health of the French Ministry of Health and the French National Research Agency published an open call for projects to support the development of digital health teaching and learning technologies, in French, and dedicated to the community of French health–related professions education institutions and professionals have free and unrestricted access to the Catalog and Index of Digital Health Teaching Resources (CIDHR) as a platform providing structured and validated information contributing to the body of knowledge necessary to master the field [9].

For example, since 1993, the URN DDH has been developing CISMeF (Catalogue et Index des Sites Médicaux en langue Française; in English, Catalog and Index of Medical Sites in French Language), a catalog of French-speaking health resources currently containing 128,689 inputs, including 9409 teaching resources. Moreover, since 1999, with the foundation of the French Medical Virtual University [10], all these teaching resources have been freely available in open access [11,12].

Dealing with teaching material in digital health for academic purposes is challenging because of the availability of many resources. However, the French-speaking material is globally limited compared with the one available in English. Therefore, we are developing the CIDHR [9].

In contrast to other educational platforms that mainly cater to English speakers and require payment, such as the Healthcare Leadership Academy [13], various platforms supported by the UK National Health Service [14], or the IMD Health cloud-based platform [15], CIDHR plays an important role in freely engaging French-speaking students and the health care practitioners community in digital health teaching and learning.

One of the primary reasons for emphasizing the need for a French-speaking knowledge catalog in the digital health domain, such as CIDHR, is to bridge the language gap. Although English
is a dominant language in scientific literature and teaching platforms, it excludes a substantial portion of the global population, particularly those more comfortable with other languages and, more particularly, French in this specific case. Thus, this language barrier can hinder the dissemination of critical information and knowledge transfer in digital health education and the development of a dedicated platform in French (which can comprise resources in other languages) [16-19].

From an informatics perspective, SaNuRN is based on semantic technologies. Since 2000, the DDH of URN has been developing and maintaining a semantic search engine (Doc’CISMeF) that was developed using primarily the Medical Subject Headings (MeSH) thesaurus [20] to manage the CISMeF resources. Starting in 2010, a multiterminology and multilingual approach is being continuously developed and used to allow any CISMeF resource to be indexed by more than 1 health terminology and by more than 1 language, although the MeSH thesaurus remains the pivotal terminology and, for CISMeF, the French and the English are the 2 pivotal languages [21,22].

As a natural evolution with the goal to share as much as possible the open access resources, and within the SaNuRN framework, starting in 2022, we have been developing and implementing a multilingual multiterminology semantic search engine CIDHR. We focus on continuously expanding CIDHR to fit the goal of the SaNuRN project and facilitating the daily teaching and learning practice in medical education by offering easy-to-use indexation and retrieval processes of any educational resource in digital health mainly toward not only French speakers but also toward others; the portal is available among other languages in English, German, Spanish, Greek, Croatian, Chinese (Mandarin), and Finish (Figure 1 [9]).

Figure 1. The Catalog and Index of Digital Health Teaching Resources (CIDHR) portal in French.

Aim, Objective, and Hypothesis

Our main aim in developing and implementing CIDHR, as a multilingual multiterminology semantic search engine, is to enhance knowledge sharing and spreading in digital health and health informatics and to expand the health-related educational community, primarily French speaking but also in other languages [23]. In particular, we aim to support the continuous development of initial (ie, bachelor level), advanced (ie, master and doctoral levels), and continuing training (ie, professionals and postgraduate levels) in digital health for health and social work fields.

Our main objective is to describe the development and implementation of the semantic search engine CIDHR in SaNuRN as a way to foster digital health education and continuous training in France. The hypothesis that guided this research is that controlled vocabularies dedicated to medical informatics and digital health, such as the Medical Informatics Multilingual Ontology (MIMO) [24,25] and the concepts structuring the FNRDH [8], to index digital health teaching and learning resources, are effectively increasing the availability and accessibility of these resources to medical students and other health care professionals.

Methods

Highlights

CIDHR is a part of the SaNuRN project. To better understand how we are developing and implementing CIDHR as a catalog of indexed digital health resources, we present the methodological steps in this process in the next lines. First, resource identification is processed by medical librarians; then, based on controlled vocabularies (an ontology and a competency referential organized as a taxonomy), the teaching and learning resources are indexed for each related knowledge domain. In the third step, the same resources are tagged with relevant academic and professional experience levels. The fourth step consists of sharing the indexed resources with the digital health teaching and learning community (with some focus on the French-speaking community). The last step consists of assessing CIDHR by obtaining informal feedback from users.

Resources Identification

To identify new or updated digital health teaching and learning resources, a group of 3 librarians from URN DDH is working on a continuous information watch, according to an internally developed and validated process comprising the steps and actions.
Thus, the librarians search proprietarily on a predefined list of academic websites of Schools of Health Sciences (eg, Medicine, Dental Medicine, Pharmacy, Nursing, Rehabilitation), National Agencies (eg, the French Ministry of Health [26]; the French National Authority for Health—La Haute Autorité de Santé [27]; the French national agency for medicines and health products safety—Agence Nationale de sécurité du médicament; and the French Agency for Food, Environmental and Occupational Health & Safety—Agence Nationale de Sécurité Sanitaire de l’Alimentation, de l’Environnement et du Travail); and other organizations involved in digital health education such as universities in France and around the world. They are also using search engine alerts, allowing reception of emails with potentially interesting content detected by their algorithms.

Moreover, the librarians monitor social media platforms, such as X (formerly known as Twitter), LinkedIn, or Facebook, by following and screening digital health–related accounts and groups sharing potentially relevant educational supports in digital health and health care informatics. The same search is performed by reading newsletters from professional organizations and academic institutions.

Furthermore, direct contacts with librarians and professional networks in digital health, particularly in the educational field, are used to obtain early updates about new and updated resources before their publishing over the web.

Resource identification also comprises the users’ engagement with CIDHR as a platform, which can share their comments with the whole team (not only the librarians) and suggest additional resources.

Therefore, by using a variety of identification approaches, the librarians involved in CIDHR can propose to the digital health experts of the SaNuRN project a wide range of digital health educational resources to integrate. It is critical to remember that the resources identified are multilingual (although mostly in French because of the SaNuRN grant requirements).

Librarians evaluate each potential resource against the following three criteria:

1. Is the resource a digital health or health informatics education–related one? The resource should be designed to teach users or to support their teaching (depending on whether the user is a student or a lecturer).
2. Is the resource accurate and up-to-date? The resource should be based on current research and best practices.
3. Is the resource accessible? The resource should be available to many users, including those with disabilities.

If a resource meets all 3 criteria, the resource is added to the SaNuRN or CIDHR repository for tagging and indexing. If a potential resource fails the evaluation, it is excluded, at least temporarily, until the librarians recheck the resource and its positive compliance with the evaluation criteria.

**Resources Indexation**

For indexing the identified educational resources, CIDHR uses 2 knowledge organization systems (KOSs).

The first is the MIMO, which comprised 3645 concepts in 33 languages as of September 2023 [23-25]. An ontology formally represents a set of concepts within a domain and the relationships between these concepts.

The second KOS was the FNRDH created in 2021. Specifically, FNRDH describes 29 different competencies and 70 different abilities. FNRDH has a 3-level hierarchy. The first relies on 5 main competencies (health data, communication in health, digital tools in health, telehealth, and cybersecurity). The second level relies on 25 subcompetencies (eg, characterizing and managing nominative data, applying [European] regulation [in particular General Data Protection Regulation]), and the last level describes 70 different abilities (eg, understanding the life cycle of the digital health data) [8].

As a side note, MIMO and FNRDH are freely available through the Health Terminology/Ontology Portal [28], also developed by URN DDH over the past 20 years [29,30]. These 2 KOSs are used at an automated stage wherein the resources are preindexed based on keyword identification and then through a librarian indexation validation stage or manual indexation if the automated process is invalid.

Moreover, CIDHR is built around 2 sets of metadata (SoM): the Learning Object Metadata (LOM) set [31] and the Dublin Core Metadata Terms (DCMI-MT) set [32]. LOM is a standard for describing digital learning resources. It provides a set of metadata elements that can be used to describe the characteristics of a learning resource such as its title, description, educational objectives, and technical requirements. DCMI-MT is a simple metadata schema that can describe various digital resources. It provides a set of 15 core metadata elements, including title, creator, and subject. Both SoM are transparent for the final user and allow efficient management of the overall available data related to a selected education resource for being included in CIDHR. These SoM are autocompleted when metadata are available with a resource (ie, a website) and are then validated by a medical librarian. If the automated process fails, the librarian handles this task.

Using 2 KOSs and 2 SoM allows a flexible and comprehensive organization of CIDHR. First, the combination of the KOSs, MIMO as an ontology, and FNRDH as a referential provides a structured way to describe the concepts and skills covered by the teaching resources. Second, the SoM provide a way to describe the characteristics of the teaching and learning resources themselves. Combining KOSs and SoM makes it easy for users to find the appropriate educational resources.

For example, a user (eg, a medical student) interested in learning about the use of artificial intelligence in digital health can use CIDHR to find learning resources that are indexed with the following MIMO concepts: “artificial intelligence,” “digital health,” “machine learning,” and “data mining”; or the same user can find resources indexed with the following FNRDH skill: “use of artificial intelligence in digital health.” Accordingly, CIDHR provides a list of relevant educational resources.
Using KOSs and metadata sets is a common practice in digital learning to organize and represent digital learning resources in a flexible, comprehensive, and user-friendly manner.

**Resources Tagging and Integration to the Curricula**

Resource indexation is a critical stage of the CIDHR knowledge management process and a pivotal component of the overall SaNuRN project. However, the main aim is to use CIDHR as a support for digital health learning and teaching in integrating the medical and health-related undergraduate, postgraduate, and life continuing education curriculum. It is also important to suggest the right resources to the specific end user (ie, student according to his degree and field of study and lecturer according to his students and his field of teaching). Thus, LOM and its instantiation in France, known as SupLomFr [33], and DCMI-MT were previously used in CISMeF that we have introduced above [11].

Thus, the 2 leading metadata are of utmost importance to help health-related students and lecturers find the right educational resources at the right time.

The first metadata is the “field of study” (ie, initial long-path education [>5 years]: medicine [Doctor of Medicine], dental surgery [Doctor of Dental Surgery], pharmacy [PharmD], and midwifery [State Diploma of Midwifery]; initial short-path education [until 5 years]: nursing [registered nurse], physiotherapy [State Diploma of Physiotherapist], and occupational therapy [State Diploma of Occupational Therapist]; and social work [State Diploma of Social Worker]).

The second metadata is the “degree level” (bachelor, master, doctorate, or residency in medicine, dental surgery, and pharmacy). It is important to point out that the graduates of an initial short-path education can continue their education in their fields at the postgraduate levels (master and doctorate degrees and lifelong continuing education).

Therefore, for any query performed on CIDHR, the end user may select and save these 2 metadata, “field of study” and “degree level” (eg, “Nursing” AND “Master Degree”; “Medicine” AND “Residency”). The so-called “training matrix” is generated to provide each combination of learners with a set of resources relevant to their profile. This set of educational resources is defined by consensus by the SaNuRN pedagogical team to be the most exhaustive. The “training matrix” is periodically updated according to the introduction of new resources or updates.

Moreover, any kind of teaching resource is cataloged in CIDHR, thanks to an extensive resources type hierarchy created for CISMeF based on a conceptual extension of the MeSH publication type [20,34]. This resource-type hierarchy has been used fruitfully for more than 20 years by users (health students, academics, and professionals) of the CISMeF platform searching for clinical-focused resources.

The following teaching resources are cataloged by tagging each one based on the following resource-type hierarchy (Figure 2 [28,35,36]): a “classical” teaching resource supporting a face-to-face course delivered with a series of slides (resource type: teaching material); evaluation of knowledge, such as multiple-choice question; and evaluation of competence, such as Objective Structured Clinical Examination or Script Concordance Test. These last 2 innovative approaches used as competency evaluation tools have been proposed for the nursing curriculum [37]; their use will be extended to other fields in CIDHR.

These combinations of the metadata tags “field of study” and “degree level” with the “resource type” tag as filters allow delivery to the user more or fewer indexed resources relevant to the knowledge fields submitted in the query to CIDHR depending on the filters selection submitted with the query.
Figure 2. List (sample) of resource types for teaching resources in the Health Terminology/Ontology Portal. CISMeF: Catalogue et Index des Sites Médicaux en langue Française (Catalog and Index of Medical Sites in French Language).

CISMeF Resources Types top tree (CISMeF resources Type)

<table>
<thead>
<tr>
<th>Description</th>
<th>Hierarchies</th>
<th>Relations</th>
<th>PubMed / Doc'CISMeF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full tree</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- CISMeF Resources Types top tree</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- audiovisual material</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- blog</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- clinical tool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- database</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- documents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- teaching material</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- audiovisual aids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- critical appraisal or critical reading</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- directed work</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- educational course</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- script concordance test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- self assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- teaching scenario</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- practicals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- problems and exercises</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- case reports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- clinical reasoning learning</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

User Experience Assessment

To assess the reception of CIDHR among users, we conducted an informal assessment including the following steps. First, a group of users consisting of both students (health students in their first year: 10/150, 6.7%) and the 5 teaching staff of digital health (JG, AB, PS, RL, and SJD) from diverse educational backgrounds and institutions was recruited. Then, immediately after the first set of lessons, the student participants were given access to CIDHR and encouraged to explore its features, search for digital health resources, and interact with the platform over a few days. Afterward, each user involved was invited to share, during a short interview, their feedback about their (1) perception of CIDHR’s user-friendliness and “easily navigable” capabilities; (2) comments on content quality comprehensiveness and the ongoing expansion; and (3) perception of CIDHR as a one-stop shop for freely and unrestricted accessible, primarily available digital health resources in their academic (ie, learning, teaching, and research) and professional activities. The last component of the feedback collection consisted of obtaining suggestions from the assessment participants.

Ethical Considerations

This research is dispensed of the ethical committee’s approval, the User Feedback for Continuous Improvement being a normal educational practice and classroom management method.
conducted in educational settings. Specifically, as non-interventional research dealing with practical habits analysis the Rouen University Hospital ethical committee does not ask for submitting such kind of research to the ethical committee. Moreover, the whole project SaNuRN that comprises CIDHR has been approved as a whole by the Delegation of Digital Health of the French Ministry of Health and the French National Research Agency [38].

Results

Resource Discovery and Indexation in CIDHR

The outcomes of the CIDHR resource identification and evaluation processes were executed by a dedicated team of 3 librarians from the URN—Rouen University Hospital DDH, aiming to collect and curate an extensive collection of digital health teaching and learning resources. Our identification strategies yielded a diverse and expansive pool of digital health educational resources through diligent exploratory searching of academic websites and platforms (e.g., a systematic review of French universities’ digital health departments and several French national agencies such as Agence Nationale de sécurité du médicament and La Haute Autorité de Santé) [26,27]. We successfully identified a continuously updating substantial number of resources catering to various aspects of digital health education. The use of search engine alerts (e.g., Google Alerts [39] and PubMed alerts [40]), social media monitoring (e.g., LinkedIn [41]), newsletters, and professional network notifications (of posts in groups of interests) also contributed significantly to the resource identification process.

In the last year, we identified approximately 500 valuable resources. It is noteworthy that the identified resources reflect a multilingual character (in particular, English). However, to align with the SaNuRN grant requirements, a substantial proportion (>90%) of the resources is in French. However, we ensured a representation of diverse languages to accommodate a wide-ranging audience interested in digital health education. The use of search engine alerts (e.g., Google Alerts [39] and PubMed alerts [40]), social media monitoring (e.g., LinkedIn [41]), newsletters, and professional network notifications (of posts in groups of interests) also contributed significantly to the resource identification process.

In addition, we supported the ongoing internationalization and a wide-ranging audience interested in digital health education. The use of search engine alerts (e.g., Google Alerts [39] and PubMed alerts [40]), social media monitoring (e.g., LinkedIn [41]), newsletters, and professional network notifications (of posts in groups of interests) also contributed significantly to the resource identification process.

In the last year, we identified approximately 500 valuable resources. It is noteworthy that the identified resources reflect a multilingual character (in particular, English). However, to align with the SaNuRN grant requirements, a substantial proportion (>90%) of the resources is in French. However, we ensured a representation of diverse languages to accommodate a wide-ranging audience interested in digital health education. The use of search engine alerts (e.g., Google Alerts [39] and PubMed alerts [40]), social media monitoring (e.g., LinkedIn [41]), newsletters, and professional network notifications (of posts in groups of interests) also contributed significantly to the resource identification process.

In the last year, we identified approximately 500 valuable resources. It is noteworthy that the identified resources reflect a multilingual character (in particular, English). However, to align with the SaNuRN grant requirements, a substantial proportion (>90%) of the resources is in French. However, we ensured a representation of diverse languages to accommodate a wide-ranging audience interested in digital health education. The use of search engine alerts (e.g., Google Alerts [39] and PubMed alerts [40]), social media monitoring (e.g., LinkedIn [41]), newsletters, and professional network notifications (of posts in groups of interests) also contributed significantly to the resource identification process.

In the last year, we identified approximately 500 valuable resources. It is noteworthy that the identified resources reflect a multilingual character (in particular, English). However, to align with the SaNuRN grant requirements, a substantial proportion (>90%) of the resources is in French. However, we ensured a representation of diverse languages to accommodate a wide-ranging audience interested in digital health education. The use of search engine alerts (e.g., Google Alerts [39] and PubMed alerts [40]), social media monitoring (e.g., LinkedIn [41]), newsletters, and professional network notifications (of posts in groups of interests) also contributed significantly to the resource identification process.

The “resource evaluation process” disclosed in the Methods section together with its 3 fundamental criteria ensures that each resource included up to now has been evaluated for relevance, “accuracy and currency,” and accessibility.

The relevance was scrutinized to ascertain its suitability for teaching and learning digital health education to serve the needs of both students and lecturers. As a result, a significant portion of the identified resources clearly aligned with digital health education objectives (323/503, 64.2%). The 35.8% (180/503) of resources that were excluded were in the scope of digital health, but they did not sufficiently focus on real teaching resources.

Furthermore, each one of the remaining resources was subjected to a rigorous assessment of “accuracy and currency” to ensure its alignment with up-to-date research findings and adherence to best practices within the digital health field. The evaluation step revealed that some resources did not meet these accuracy and currency criteria and were rejected (approximately 36%).

The “accessibility” of the educational supports is a critical aspect emphasized in CIDHR resource evaluation to include in the catalog materials that can effectively be used by a broad range of the digital health educational community, including individuals with disabilities. This evaluation highlighted the commitment of many resources to accessibility.

If a potential resource does meet any one of these criteria, it does not move to inclusion in CIDHR and remains in a secondary list of resources to be periodically re-evaluated for future inclusion.

Resources that successfully passed all 3 evaluation criteria were promptly included in CIDHR. These resources are diligently indexed and tagged as described in the Methods section.

Tailored Learning Paths: Metadata, Training Matrix, and Resource Cataloging in CIDRH

The semantic search engine of CIDHR based on MIMO and FNRDH allows user-friendly access to previously indexed and tagged resources. At the end of September 2023, CIDHR comprised 371 available resources in the digital health field relevant to students and teaching staff from the first academic year of academic studies to lifelong continuing education. The French grant required that 80% of the effort should focus on the bachelor “degree level.” Therefore, approximately all the 371 resources included in CIDHR are focusing on bachelor’s students.

CIDHR is constantly expanding, with plans to incorporate increasingly as much as possible digital health teaching resources from the French health–related studies curricula over the next few years [6].

Figure 3 shows an example of the results for the query “dossiers médicaux électroniques” (in English, “electronic health records” or EHRs).
Figure 3. Example of results to the query “dossiers médicaux électroniques” (in English, “electronic health records” or EHRs). CIDHR: Catalog and Index of Digital Health Teaching Resources; CISMeF: Catalogue et Index des Sites Médicaux en langue Française (Catalog and Index of Medical Sites in French Language).

Figure 4 shows an example of a digital health educational resource, as a bibliography card, indexed using MIMO and FNRDH, which is an example of CIDHR’s capabilities. A CIDHR bibliographic card comprises the following metadata: (1) the resource title, (2) the resource publisher or author, (3) the country of the source, (4) the year of publication, (5) the type of resource, (6) an abstract presenting the resource, and (7) a list of the terms and concepts used to index the resource with regard to controlled vocabularies and referential such as MIMO and FNRDH (Figure 4).

Figure 4. Example of an indexed resource in Catalog and Index of Digital Health Teaching Resources (CIDHR) comprising the following metadata: resource title, resource published and author, country, year of publication, type of document, an abstract, and a list of the terms and concepts used for indexation (here with both Medical Informatics Multilingual Ontology [MIMO] and French National Referential on Digital Health [FNRDH]).

The resource is written in French and focuses on EHRs, a concept defined in both MIMO (ie, “dossiers médicaux électroniques”) and FNRDH (ie, “Interagir de manière adaptée entre professionnels, avec l’usager, les aidants et accompagnants et avec les institutions et administrations,” in English, “Interact in an appropriate manner between professionals, with the healthcare customer, caregivers and companions and with institutions and administrations”; and “Utiliser les outils et services sociaux adaptés et identifier leur articulation avec d’autres dossiers partagés,” in English, “Use the appropriate basic tools and services and identify their connection with other shared files”). It educates the learners on the fundamentals and the importance of the EHRs, making it an invaluable resource for anyone looking to enhance their digital health knowledge.

To facilitate the indexing process with FNRDH, which presents considerable complexity for medical librarians, the SaNuRN pedagogical team has established manual associations between
MIMO and FNRDH concepts. For instance, this involves manually linking the MIMO concept with the FNRDH competency. It is essential to clarify that this mapping relation does not constitute a strict “exact match”; instead, it means that when a librarian indexes a teaching resource using a MIMO concept (eg, “electronic medical records”) associated with an FNRDH ability (eg, “Interact appropriately between professionals, with the healthcare customer, caregivers and companions and with institutions and administrations”), the educational resource is also indexed with this corresponding FNRDH competency.

Nevertheless, certain cases require manual indexing with FNRDH by medical librarians, primarily because of the absence of the MIMO concepts for specific capacities, still not defined and implemented in MIMO, such as the “lifecycle of health data.” Thus, to minimize the dependency on manual FNRDH indexing, the SaNuRN pedagogical team is actively developing MIMO concepts and establishing mappings between MIMO and FNRDH concepts, including those pertaining to the lifecycle of health data.

In addition, as a part of CIDHR capacities, the end-user process for any query to deliver an organized list of educational resources is considered. The first item on the list must be studied first, followed by the second item, and so on. This organized list is manually created for each FNRDH competency; in other words, we create a breadcrumb navigation for teaching and learning resources linked to each FNRDH competency. Currently, this organized list is familiar to all the students in all the fields of study. In the future, this organized list will be, when relevant, adapted to fit with the requirements of each field of study (eg, medicine, nursing), degree level (eg, bachelor, residency), and targeted level competencies or skills (eg, beginner, intermediate, and advanced).

User Feedback for Continuous Improvement
To assess CIDHR’s usability and acceptance among users, we collected informal feedback from a select group comprising both first-year health students (10/15, 67%) and teaching staff (5/15, 33%). Their feedback universally reflected a positive sentiment, characterizing the platform as remarkably user-friendly and easily navigable. Moreover, they lauded the platform’s existing resource collection, founded on rigorous content quality control, and appreciated its ongoing expansion. Notably, users articulated their assessment, highlighting CIDHR’s comprehensiveness, precision, and user-friendliness. Nonetheless, their constructive suggestions included the need for augmenting multilingual resources and offering more comprehensive resource information, particularly with respect to metadata. In the users’ collective perception, CIDHR was deemed a one-stop destination for discovering high-quality digital health resources. An additional commendable attribute was the platform’s unrestricted accessibility, which rendered it a valuable asset for all users.

Moreover, additional suggestions related to the need for more multilingual resources and comprehensive metadata were noted (eg, field of study, resource language, and resource scoring; Table 1).

Table 1. Summary of the feedback collected during the Catalog and Index of Digital Health Teaching Resources user experience informal assessment.

<table>
<thead>
<tr>
<th>Feedback category</th>
<th>Students</th>
<th>Lecturers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usability</td>
<td>User-friendliness and “easily navigable” capabilities</td>
<td>User-friendly and “simple to understand”</td>
</tr>
<tr>
<td>Content quality</td>
<td>Valuable, “easy to understand”</td>
<td>Valuable, comprehensiveness</td>
</tr>
<tr>
<td>One-stop shop potential</td>
<td>Free resources, easy to access, on various relevant content</td>
<td>Real one-stop shop freely and unrestricted accessible, especially available digital health resources in their academic (ie, learning, teaching, and research) and professional activities</td>
</tr>
<tr>
<td>Participants suggestions for improvements</td>
<td>More than French-only resources, in particular English, but also Arabic, Spanish and Portuguese (native language of the students)</td>
<td>More metadata on bibliographic card; more than French-only resources, in particular English</td>
</tr>
</tbody>
</table>

Discussion

Overview
The integration of digital technologies in health care and medical education is becoming increasingly vital. This study introduces CIDHR as part of the SaNuRN project to enhance digital health education in France. CIDHR is a comprehensive digital platform that indexes and organizes educational resources related to digital health, catering to students and health care professionals. This discussion explores the strengths and limitations of CIDHR, potential future perspectives, and the impact on digital health education.

Strengths and Limitations
CIDHR is the heart of a digital health educational platform that provides an extensive array of inclusive and accessible teaching and learning resources to a diverse global audience in the health care professional education landscape. CIDHR has a large and continuously expanding collection of up-to-date and relevant digital health resources that serves as a one-stop shop related to all aspects of digital health education needs, catering to lecturers, students, and professionals alike.

CIDHR is committed to providing comprehensive support to French-speaking individuals seeking digital health education. To ensure that language barriers do not impede access to educational resources, CIDHR has indexed a wide range of materials in multiple languages, in addition to its French language resources. These materials are designed to cater to

https://mededu.jmir.org/2024/11/e48393

(Manuscript submitted 16 October 2021, revised 17 January 2023, accepted 28 January 2023)
diverse linguistic needs and are available to all individuals seeking to enhance their digital health knowledge. With CIDHR’s vast collection of indexed educational resources, individuals can access high-quality information and support regardless of their native or daily spoken language.

To improve resource indexing and search precision, CIDHR uses controlled vocabularies such as MIMO and FNRDH, which enable users to locate relevant educational materials that align with their specific digital health skills and competencies with ease. Moreover, CIDHR prioritizes resource accessibility, making its platform suitable for a broad audience, including individuals with disabilities [42,43]. Thus, CIDHR, being based on a multilingual semantic search engine, would enhance accessibility and inclusivity. By looking at all (even mainly French speakers currently) health care professionals, researchers, and students, CIDHR allows them to have access to a broader range of educational resources, fostering a more inclusive learning environment. This inclusivity aligns with the principles of health equity and diversity in medical education [44]. Furthermore, the CIDHR platform’s user-friendly interface and straightforward navigation enable users to connect with relevant educational resources quickly and efficiently.

By looking at these advantages and the SaNuRN aim to facilitate digital health educational resources, the current corpus, including 371 elements, will be expanded by continuing the collection and evaluation process, in parallel with cooperation with as many possible faculties and schools of health (ie, 31 medical schools in France). We expect approximately 700 CIDHR resources by mid-2024.

However, some limitations have been identified. First, although CIDHR supports mainly French resources, it would benefit from expanding its multilingual and international support to make it more accessible to a global audience of the digital health education community. Second, it is necessary to expand CIDHR resource collection to incorporate more digital health resources from diverse sources allowing providing them to the educational community and industry insights. Third, although SaNuRN plans to provide personalized learning paths to users, via CIDHR, it is crucial to ensure that these paths are effective and tailored to the individual needs of each user, which requires further research and development [45,46]. Fourth, integrating CIDHR with the learning management systems used by educational institutions would streamline access to digital health resources for students and educators (eg, Moodle [48]). Finally, CIDHR will benefit from the development of a feedback and rating system for resources not only to help users identify the most valuable and reliable materials within the platform but also to allow the SaNuRN team project to get feedback on the resource collection, indexing, and tagging processes from mass users’ practice. All these measures will augment CIDHR utility and enrich the user experience.

**Conclusions**

CIDHR represents a significant advancement in digital health education, offering a diverse, accessible, and validated resource collection. Although it has strengths in its multilingual approach, controlled vocabularies, and user-friendliness, addressing resource evaluation challenges and enhancing resource information are areas for continuous improvement. The future perspectives for CIDHR include further expansion, collaboration, personalized learning, integration, and user feedback mechanisms, all aimed at enriching the digital health education experience for students and health care professionals.

To the best of our knowledge, no prior published research has described a multilingual semantic search engine to query a digital health educational repository to be used by any health-related field student and lecturer. This is also because of the uniqueness of the development of the Health Terminology/Ontology Portal and MIMO by the members of the SaNuRN team. These projects have no equivalent to date.

The hypothesis that guided this part of the SaNuRN research and that we have validated is that controlled vocabularies and knowledge and skills referential dedicated to medical informatics and digital health, such as MIMO [22,23] and FNRDH [24], to index related educational resources, are effectively increasing the availability and accessibility of these resources to the health care–related community. This approach is possible as MIMO and CIDHR search engine are multilingual.

A European project called the HosmartAI (Hospital Smart AI) project deals with the digital transformation of the European health care sector to make the European health care system more strong, efficient, sustainable, and resilient. CIDHR can play an important role in acquisition of literacy in digital health for professionals [49]. The European
Federation for Medical Informatics is taking part in different projects such as HosmartAI and as a collaboration and cooperation-oriented scientific and academic international organization, it can help disseminate information about CIDHR to promote its use by an increasing number of members of the digital health educational community worldwide.

However, the need to develop and improve digital health competencies for medical learners and broadly for health-related students and professionals is an established objective worldwide [45,50,51]. As a fact, prior studies evaluating digital health competencies among German medical students have shown a significant improvement after a digital health teaching course was introduced in their curriculum, although most students found that digital health is not sufficiently taught in undergraduate medical education, while it may influence everyday work of physicians [52].

Thus, CIDHR will have an important role on the educational grounds to improve digital health literacy of students and lecturers and to increase their engagement with these ubiquitous ways of delivering and receiving health care [46,53].

CIDHR is a fair and findability, accessibility, interoperability, and reusability principles–focused platform looking at making “findable” educational resources by using a one-stop-shop portal approach, “accessible” by integrating these resources available overtime and by anyone (ie, including people with disabilities), “interoperable” by making these resources readable in the most common formats (PDF files and video and audio support on browser-embedded readers, such as YouTube), and finally “reusable” by providing resources freely distributed and under open access licensing [54-56].

Acknowledgments
This work was partially supported by the SaNuRN project (ANR_22-CMAS-0014 3,951,200) granted by the Delegation of Digital Health of the French Ministry of Health and the French National Research Agency.

This work was partially supported by the HosmartAI (Hospital Smart development based on AI) project that received funding from the European Union’s Horizon 2020 research and innovation program under grant #101016834.

Data Availability
The data sets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Authors’ Contributions
JG was involved in conceptualization, methodology, software, validation, formal analysis, and investigation and prepared the original draft and reviewed and edited the manuscript. AB contributed to methodology, validation, formal analysis, investigation, and reviewing and editing the draft and acquired funding (HosmartAI [Hospital Smart development based on AI]). F Dufour was involved in validation and reviewed and edited the draft. F Disson performed the software analysis. BD was involved in methodology and software analysis. HC was involved in project administration. RL was involved in conceptualization, methodology, and reviewing and editing the draft. MF reviewed and edited the draft and supervised the study. PS was involved in conceptualization, methodology, formal analysis, and resources; reviewed and edited the draft; supervised the study; and acquired funding (SaNuRN [Santé Numérique Rouen Nice]). SJD was involved in conceptualization, methodology, validation, formal analysis, and resources; prepared the original draft and reviewed and edited the manuscript; supervised the study; participated in project administration; and acquired funding (SaNuRN).

Conflicts of Interest
None declared.

References
1. Steinhubl SR, Topol EJ. Digital medicine, on its way to being just plain medicine. NPJ Digit Med 2018 Jan 15;1(1):20175 [FREE Full text] [doi: 10.1038/s41746-017-0005-1] [Medline: 31304349]


15. IMD Health home page. IMD Health. URL: https://www.imdhealth.com/ [accessed 2024-01-04]


32. DCMI metadata terms. DublinCore. URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ [accessed 2024-01-04]


41. LinkedIn log in. LinkedIn. URL: https://www.linkedin.com/ [accessed 2024-01-04]


Abbreviations

CAU: Côte d’Azur University
CIDHR: Catalog and Index of Digital Health Teaching Resources
CISMeF: Catalogue et Index des Sites Médicaux en langue Française (Catalog and Index of Medical Sites in French Language)
DCMI-MT: Dublin Core Metadata Terms
Unpacking the Experiences of Health Care Professionals About the Web-Based Building Resilience At Work Program During the COVID-19 Pandemic: Framework Analysis

Wei How Darryl Ang¹, RN, BSN (Hons), PhD; Zhi Qi Grace Lim¹, BA; Siew Tiang Lau¹, RN, BHS, MHS, PhD; Jie Dong¹, RN; Ying Lau², RN, RM, BSc, BN (Hons), MN, PhD

¹Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
²The Nethersole School of Nursing, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Ying Lau, RN, RM, BSc, BN (Hons), MN, PhD
The Nethersole School of Nursing, Faculty of Medicine, The Chinese University of Hong Kong
Room 829, 8/F, Esther Lee Building, The Chinese University of Hong Kong, Shatin, New Territories
Hong Kong
China (Hong Kong)
Phone: 852 39436222
Fax: 852 26035269
Email: yinglau@cuhk.edu.hk

Abstract

Background: The COVID-19 pandemic has resulted in a greater workload in the health care system. Therefore, health care professionals (HCPs) continue to experience high levels of stress, resulting in mental health disorders. From a preventive perspective, building resilience has been associated with reduced stress and mental health disorders and promotes HCPs’ intent to stay. Despite the benefits of resilience training, few studies provided an in-depth understanding of the contextual factors, implementation, and mechanisms of impact that influences the sustainability of resilience programs. Therefore, examining target users’ experiences of the resilience program is important. This will provide meaningful information to refine and improve future resilience programs.

Objective: This qualitative study aims to explore HCPs’ experiences of participating in the web-based Building Resilience At Work (BRAW) program. In particular, this study aims to explore the contextual and implementational factors that would influence participants’ interaction and outcome from the program.

Methods: A descriptive qualitative approach using individual semistructured Zoom interviews was conducted with participants of the web-based resilience program. A framework analysis was conducted, and it is guided by the process evaluation framework.

Results: A total of 33 HCPs participated in this qualitative study. Three themes depicting participants’ experiences, interactions, and impacts from the BRAW program were elucidated from the framework analysis: learning from web-based tools, interacting with the BRAW program, and promoting participants’ workforce readiness.

Conclusions: Findings show that a web-based asynchronous and self-paced resilience program is an acceptable and feasible approach for HCPs. The program also led to encouraging findings on participants’ resilience, intent to stay, and employability. However, continued refinements in the components of the web-based resilience program should be carried out to ensure the sustainability of this intervention.

Trial Registration: ClinicalTrials.gov NCT05130879; https://clinicaltrials.gov/ct2/show/NCT05130879

doi:10.2196/49551

KEYWORDS
resilience; intent to stay; employability; health care professionals; process evaluation; framework analysis; framework; resilience; stress; mental health disorder; prevention; training; qualitative study; web-based tool; tool; sustainability
Introduction

Background

The emergence of the COVID-19 pandemic has led to extensive changes in the health care landscape. Globally, the repeated waves of COVID-19 infections have led to health care professionals (HCPs) grappling with occupational health hazards and overstretched assignments [1,2]. These constant stressors have led to HCPs experiencing a surge in symptoms of burnout, insomnia, and mental health distress [3-5]. Accordingly, the intensification of physical and mental exhaustion has led to a considerable increase in the turnover of HCPs [6]. With a smaller health care workforce, health care administrators need to prioritize and concentrate their efforts on enforcing supportive measures to ensure that HCPs continue to be inoculated against stress and mental health disorders. Thus, reducing workplace-related stress may have encouraging effects on HCPs’ intent to stay [7,8].

Contemporarily, more persuasive evidence has alluded to the importance of noncognitive skills as protective factors against mental health distress [9,10]. An emerging interest among noncognitive skills is the development of an individual’s resilience. Resilience is the ability to overcome adversities [11,12]. Theoretically, resilience can be understood from various perspectives, as a trait (eg, personality), process (eg, interaction with protective factors), or outcome (eg, becoming resilient). More importantly, building an individual’s resilience has positive effects on their mental well-being [13,14].

Figure 1. Process evaluation framework.

![Process Evaluation Framework](image)

First, contextual factors are unique situational factors that influence how the intervention may be delivered or have affected the participants [25]. These contextual factors may have eventual implications on the implementation and mechanisms of impact. Second, the implementation process is the identification of factors that may influence the delivery of the intervention [25]. This may include the collection of data that reflects intervention fidelity [26]. Third, mechanisms of impact describe participants’ responses to and interaction with the intervention. In addition, mechanisms of impact identify any potential mediators, pathways, or consequences as a result of their participation in the intervention [25]. Thus, conducting process evaluations of interventions may be worthy in providing recommendations for improvements and supporting the eventual implementation of the program. Although prior qualitative evaluations of resilience programs [22,27,28] have made valuable contributions toward an in-depth understanding of participants’ experiences, its findings may not be transferrable because of several factors, such as population, cultural differences, and type of resilience program. For these reasons, conducting a study to encapsulate the experiences of the participants of the Building Resilience At Work (BRAW) program is important.

Objectives

This qualitative study explores HCPs’ experiences of participating in the BRAW program. Guided by the process evaluation framework [25], this study also aims to examine the contextual and implementation factors that affected participants’ experiences and identify the outcomes that arose from their participation in the BRAW program.
Methods

Ethical Considerations
This study was approved by the National University of Singapore Institutional Review Board (NUS-IRB-2021-703). This study’s procedures were followed in accordance with the Declaration of Helsinki. Eligible participants were recruited from August 2021 to December 2022. Participants were provided with a participation information sheet, and they were allowed to withdraw without penalty. After obtaining informed consent, participants were invited to participate in a web-based semistructured audio- and video-recorded interview via Zoom (Zoom Video Communications). The interview transcripts were de-identified and coded using pseudonyms. Participants were given 20 Singapore Dollars for completing the study.

Research Design
This qualitative study is part of a randomized controlled study conducted in Singapore (ClinicalTrials.gov NCT05130879). A process evaluation approach [25] comprising semistructured individual digital interviews was undertaken to explore participants’ experiences of using the web-based BRAW program. This study is reported based on the COREQ (Consolidated Criteria for Reporting Qualitative Research) [29] (Multimedia Appendix 1).

Setting and Participants
This study was conducted from April 2021 to December 2022 in Singapore, a multiethnic and multicultural city-state. Based on the national census [30], there are approximately 70,178 registered HCPs, and most of them are nurses (61.27%). Participants were eligible to participate in this qualitative study if they were practicing as an HCP in Singapore, could comprehend the English language, had access to a device that could connect to the internet, and completed the web-based BRAW program. A total of 33 participants who completed the web-based BRAW program were purposively sampled to participate in this qualitative study.

Web-Based BRAW Program
The web-based BRAW program is a 6-session weekly web-based program hosted via Microsoft Teams (Microsoft Corp). The resilience program was developed based on a systematic review [13] and evidence-based therapies, such as cognitive behavioral therapy [31], acceptance and commitment therapy [32], and problem-solving model [33]. The BRAW program comprised 6 different topics, namely, happiness and positivity, cognitive restructuring, behavioral activation, emotion regulation, positive work climate, and problem-solving (Table 1). It also comprised several elements, short videos, quizzes, and homework (Figure 2). A web-based forum was also provided for participants to interact with each other and provide social support.

Figure 2. Elements of the web-based BRAW program. BRAW: Building Resilience At Work.
Table 1. Overview of the Building Resilience At Work program.

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Happiness and positivity</td>
<td>• Understanding strengths and resilience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Fostering positive attitude</td>
</tr>
<tr>
<td>2</td>
<td>Cognitive restructuring</td>
<td>• Identifying dysfunctional automatic thoughts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Using cognitive behavioral techniques to modify dysfunctional thoughts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Formulating rational responses to automatic thoughts</td>
</tr>
<tr>
<td>3</td>
<td>Behavioral activation</td>
<td>• Initiating and using behavioral activation techniques</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Building healthy interpersonal relationships and peer support</td>
</tr>
<tr>
<td>4</td>
<td>Emotion regulation</td>
<td>• Regulating emotions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Preventing and managing conflict</td>
</tr>
<tr>
<td>5</td>
<td>Positive work climate</td>
<td>• Forging a supportive work environment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Developing supportive collegial relationships</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Promoting coworker support</td>
</tr>
<tr>
<td>6</td>
<td>Problem-solving</td>
<td>• Solving work-life problems using a framework</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Importance of work-life balance</td>
</tr>
</tbody>
</table>

Data Collection

The digital interviews were scheduled at a time convenient for the participants. Participants were reminded to ensure that their cameras and microphones were working prior to the interviews. All interviews were conducted by a female researcher (ZQGL) who received formal training in qualitative research. The interviewer was supported by 2 doctoral-prepared researchers (WHDA and YL) who are experienced in qualitative research. During the digital interview, the interviewer started by building rapport with the participants and sharing the aims and processes of this study. In addition, sociodemographic characteristics including age, sex, ethnicity, and occupation were collected. Afterward, the interview was conducted according to the semistructured guide. The guide was developed based on the process evaluation framework [25] and comprised open-ended questions. Then, the initial guide was circulated to the research team and refined. Subsequently, the interview guide was piloted among 5 participants and was further revised for clarity. The final interview guide can be found in Textbox 1. The mean duration of the interviews was 35.48 (SD 7.83; range 20-54) minutes. Data saturation was achieved at the 31st participant, and 2 additional interviews were conducted to confirm saturation [34].
Textbox 1. Semistructured interview guide.

<table>
<thead>
<tr>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What was your experience when completing the Building Resilience At Work (BRAW) training program?</td>
</tr>
<tr>
<td>2. What were the issues with the platforms for the training sessions that you have encountered?</td>
</tr>
<tr>
<td>3. How did you feel about the duration of each training video?</td>
</tr>
<tr>
<td>4. How did you feel about the quizzes?</td>
</tr>
<tr>
<td>5. How did you feel about the homework?</td>
</tr>
<tr>
<td>6. How did you feel about the forum?</td>
</tr>
<tr>
<td>7. How did you feel about the entire duration of the 6-week BRAW training program?</td>
</tr>
<tr>
<td>8. What were the aspects of the intervention (eg, homework, quizzes, and forum) that you particularly liked or disliked?</td>
</tr>
<tr>
<td>9. Were there any sessions that stood out?</td>
</tr>
<tr>
<td>10. How did you feel about the contents?</td>
</tr>
<tr>
<td>11. Could you tell me your overall experience with applying the strategies learned from the BRAW intervention at work?</td>
</tr>
<tr>
<td>12. How was your experience of applying the strategies at work?</td>
</tr>
<tr>
<td>13. Did you encounter any problems or frustrations when trying to apply the strategies at work?</td>
</tr>
<tr>
<td>14. Has the BRAW training program influenced your resilience at work?</td>
</tr>
<tr>
<td>15. Has the BRAW training program influenced your enthusiasm and dedication at work?</td>
</tr>
<tr>
<td>16. Has the BRAW training program influenced your intention to leave?</td>
</tr>
<tr>
<td>17. Has the BRAW training program influenced your ability to gain and maintain employment?</td>
</tr>
<tr>
<td>18. Has the BRAW training program influenced your work performance?</td>
</tr>
<tr>
<td>19. Are there any other strategies that would help you to manage stress and build resilience that we have not mentioned in the BRAW intervention?</td>
</tr>
<tr>
<td>20. Do you have anything else to add that we have not covered in this interview?</td>
</tr>
<tr>
<td>21. Finally, are you okay for me to contact you for some follow-up questions?</td>
</tr>
</tbody>
</table>

Data Analysis

The video-recorded interviews were transcribed verbatim by 1 researcher (ZQGL) and verified for accuracy by another researcher (WHDA). The transcripts were imported and analyzed using NVivo (version 12; Lumivero). Transcripts were returned to the participants for their comments. A deductive framework analysis method [35] was then undertaken as it provides a systematic approach to analyzing qualitative data [36]. In addition, the use of a matrix structure provides a visually straightforward recognition of patterns in the data that can be useful in identifying similarities or differences between participants’ narratives [36]. In line with the research questions, a framework analysis approach is suitable, as this study was guided by the process evaluation framework and sought to examine participants’ experiences of the BRAW program. Particularly, it identifies the contextual and implementation factors that affected their participation and the outcomes of participation.

A 5-step framework analysis approach [35,37] was independently performed by 2 researchers (WHDA and YL). First, the researchers familiarized themselves with the data by reading the transcripts accompanied by listening to the interviews. Second, the transcripts were coded based on the process evaluation framework [25]. After completing the coding for the first 5 transcripts, both researchers compared their codes and developed a standardized code book. Following discussions among the researchers, the eventual code book comprised 11 different categories.

Third, after completing the coding for all transcripts, a total of 347 codes were brought together and discussed among the researchers. The similarities and differences that arose during the coding process were deliberated. Cohen κ was used to calculate the interrater agreement for the coding, and good agreement was found (κ=0.79). Consequently, the codes were organized and indexed based on the process evaluation framework. Fourth, the codes were further reduced by summarizing the key information for the indexed data in each category. Finally, the identified codes were mapped using a coding tree (Table S1 Multimedia Appendix 2) and interpreted using visual and narrative forms. Finally, 3 themes and 7 subthemes were derived from the framework analysis. The themes and subthemes were provided to a select group of participants who were willing to provide feedback on the findings.

Rigor

The principles of credibility, transferability, dependability, and conformability were used to demonstrate rigor [38]. First, a reflexivity journal was maintained by all members of the research team to improve their self-awareness and reduce any potential personal influences on the data. Second, the data
analyses were conducted by 2 independent researchers (WHDA and YL). Third, participants were invited to review their transcripts to clarify the context of the statements and ensure that the final themes and subthemes were representative of their experiences [39]. Subsequently, an audit trail detailing the recruitment, data collection, and analysis process was conducted to ensure ease of replication, transparency, and dependability [38]. Finally, a thick description of the context and the intervention was provided, this facilitates the transferability of the findings of this study [38].

### Results

#### Overview

A total of 33 HCPs participated in this qualitative study. The sociodemographic variables are presented in Table 2. Most of the participants were between the ages of 31-40 years (n=11, 34%), female (n=24, 73%), ethnic Chinese (n=25, 76%), and nurses (n=15, 46%). The findings from the framework analysis unveiled 3 themes and 7 subthemes that depicted participants’ experiences, interactions, and impacts from the BRAW program. The 3 themes were learning from web-based tools, interacting with the BRAW program, and promoting participants’ workforce readiness (Figure 3).

#### Table 2. Participants sociodemographic characteristics (N=33).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age group (years), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>21-25</td>
<td>5 (15)</td>
</tr>
<tr>
<td>26-30</td>
<td>9 (27)</td>
</tr>
<tr>
<td>31-40</td>
<td>11 (34)</td>
</tr>
<tr>
<td>41-50</td>
<td>6 (18)</td>
</tr>
<tr>
<td>51-60</td>
<td>2 (6)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>9 (27)</td>
</tr>
<tr>
<td>Female</td>
<td>24 (73)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>25 (76)</td>
</tr>
<tr>
<td>Malay</td>
<td>7 (21)</td>
</tr>
<tr>
<td>Indian</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Profession, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Allied health worker</td>
<td>12 (36)</td>
</tr>
<tr>
<td>Clinical administrator</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Clinical researcher</td>
<td>4 (12)</td>
</tr>
<tr>
<td>Nurse (registered and enrolled)</td>
<td>15 (46)</td>
</tr>
<tr>
<td>Physician</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Duration of interviews (minutes)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>35.48 (7.83)</td>
</tr>
<tr>
<td>Range</td>
<td>20-54</td>
</tr>
</tbody>
</table>
Figure 3. Participants’ experiences of the web-based BRAW program. BRAW: Building Resilience At Work.

Theme 1: Learning From Web-Based Tools

Overview
The first theme depicts the BRAW implementation process. It particularly describes how participants learned through web-based tools via Microsoft Teams. This is elaborated in 2 subthemes, namely, engaging with web materials and internalizing the resilience process.

Engaging With Web Materials
The BRAW program provided various web materials, ranging from short videos to quizzes and homework. The short videos were developed using animations, graphics, and subtitles, which appealed to the participants and supported their engagement with the web materials:

The use of graphics was quite good, the animations and all, so like, it kept me wanting to finish watching, not like stop halfway. Yeah...the pace was also good, and like, just nice, not too much information overload.
[Participant 24, female, Chinese, nurse]

However, some participants were encumbered by the number of tasks (eg, weekly quizzes and homework). For instance, the weekly homework was described to be a “chore,” and this can be a disincentivizing factor in completing the program. As an alternative, a participant proposed that renaming the weekly tasks could be a strategy to overcome the inertia:

Because “homework” it sounds like “ task,” erm, like a chore to be done, you know, but “reflection” is like, you reflect on what you-you-you need to do. So, sounds more forgiving. [Participant 26, female, Malay, nurse]

Internalizing the Resilience Process
Despite the conflicting work commitments and activities in the BRAW program that participants had to undergo, they credited the quizzes and homework as factors that supported the internalization of the learning process. Particularly, reviewing the questions found in the quizzes and homework facilitated an internalization process:

Just by plain reading the question, it may set you thinking, you see. You don’t know what’s happening or your subconscious, you’re already motivated right, you learn some new content. And that homework may actually be building synapses, you know, trying at the backend that you don’t know about. [Participant 10, male, Malay, physician]

However, not all participants were well-versed in the contents of the BRAW program. Several participants highlighted difficulties in appreciating the theoretical aspects of the program:

When it gets a little bit more “science-y,” like the brain and then they tell you, I don’t know all the words, I don’t remember, but like the brain and then, certain kinds of thoughts and all that. Then, those kinds of stuff, no, like I haven’t heard of that before. [Participant 15, female, Indian, clinical researcher]

Notwithstanding, these groups of participants, particularly those who did not receive formal training in health sciences, verbalized how they used the quizzes as an avenue to understand the various technical terms that they were not familiar with:

Especially some of the terms, erm, maybe a bit technical? I’m not that acquainted. So, it referring to the quizzes allows me to clarify, review and
understand and get it correct. [Participant 8, female, Chinese, clinical administrator]

Theme 2: Interacting With the BRAW Program

Overview
The second theme describes the BRAW program’s mechanism of impact and the relevant contextual factors that influenced it. This theme expressed how participants responded and interacted with the BRAW program and is highlighted in 2 subthemes, namely, appreciating the asynchronous self-paced program and relating to the applicability of the contents.

Appreciating the Asynchronous Self-Paced Program
Due to the higher workload brought upon by the COVID-19 pandemic and the resumption of usual clinical duties, participants had to contend with numerous conflicting priorities. Hence, they appreciated how the BRAW program was designed as an asynchronous self-paced program. This allowed them to learn at their own pace and time:

Healthcare workers are busy, so they don’t have to find a specific day and time to attend an intervention, whether be it online or on-site, face-to-face or whatever, so having something that you can access on your own time and target is good. [Participant 4, female, Chinese, clinical researcher]

However, despite the self-paced nature of the program, participants struggled with finding suitable time outside their personal commitments and rest to engage in the program. This was more prominent among HCPs who are on shift work duties:

We are really packed and rushed at work, and there’s a lot of multitasking. It’s like very draining at work. I think the shifts also, so you do rotating shifts. So, it’s quite tiring after work to find time. [Participant 5, female, Chinese, nurse]

Nevertheless, some participants felt that introducing more web-based synchronous elements through videoconferencing tools may be able to better support their learning:

These sessions were to be interactive whereby we can do it via Zoom, to share every participant’s experience, it would be even better. [Participant 28, female, Chinese, nurse]

Relating to the Applicability of the Contents
The BRAW program was conducted at the peak of the COVID-19 pandemic in Singapore. Due to the stressors inflicted by the additional workload, participants felt that the program was delivered at an opportunistic time to support their psychological well-being:

I think you kind of met me at the right time and I feel that I need to self-improve. [Participant 3, male, Chinese, nurse]

In particular, participants appreciated how the contents were relatable to their concerns and felt that they were able to translate their newly acquired theoretical knowledge to an actual situation:

I really appreciate the teamwork and emotional regulation, like the ones I could really practice, putting time for myself, things like that. [Participant 6, male, Chinese, nurse]

Theme 3: Promoting Participants’ Workforce Readiness

Overview
The final theme describes how the BRAW program has influenced participants’ readiness to maintain in the workforce. Through participants’ narratives, the BRAW program has a profound impact on their resilience, intent to stay, and employability. This theme is further elaborated in 3 subthemes, namely, drawing resilience in times of adversity, promoting intent to stay in health care, and becoming future-ready.

Drawing Resilience in Times of Adversity
The BRAW program instilled numerous positive aspects in participants. As participants translated their newly acquired knowledge into practice, they demonstrated resiliency by overcoming the challenges and difficulties experienced in the workplace:

Yup, especially when dealing with negative emotions and how to bounce back up again. [Participant 1, male, Chinese, nurse]

When asked about the extent of the improvements, the majority of the participants felt noticeable improvements. For instance, they observed an evident increase in their ability to overcome situations:

In the past...I take quite a while to recover...Then, nowadays, it’s a bit better, even though I think about it, I can move on from it. And I can have a more positive mindset about it. So, I don’t blame myself for something that happened, or I don’t dwell on the thing that happened. Instead, I focused on the future, like if it happens again, what can I do. [Participant 13, female, Chinese, audiologist]

Promoting Intent to Stay in Health Care
Participants also felt that the BRAW program supported their resilience to remain steadfast in the health care sector. This was an interesting viewpoint expressed by most participants because it proposes that the improvement of psychological well-being has increased their intent to stay in their current role:

This course [referring to the BRAW program] actually helps me dispel away negative thoughts, put things in perspective, and reframe my mind away so that I can still go through the job. [Participant 14, female, Malay, medical technician]

However, most of the participants also felt that resilience training alone may not be sufficient to influence their intent to stay. Instead, one’s intent to stay may be influenced by a larger environmental factor such as management-related reasons:

The management did not do anything, so I feel that I should just quit this organization because they don’t take care of us. [Participant 25, male, Malay, nurse]
Becoming Future-Ready

The majority of the participants felt that resilience is a form of a positive attribute. When asked if being resilient is an important factor in securing employment, participants felt that resiliency was a personal competency and may have indirect impacts on getting one employed:

I won’t say, it’s directly, okay, this [referring to the BRAW program] will help you get the job, but it’s more of like okay, it helps you work on yourself as a person. So, that indirectly translates to being a more employable person. [Participant 13, female, Chinese, audiologist]

Nevertheless, participants perceived that the contents of the BRAW program could help shape an individual’s emotional quotient. This may translate to the development of one’s leadership skills:

It [referring to the BRAW program] shapes a person who has a lot of EQ and understanding...So, I think it does make, if you can master these techniques very well, I do believe that it can make you a better leader. [Participant 12, male, Chinese, respiratory therapist]

Discussion

Principal Findings

This qualitative study aimed to explore HCPs’ experiences of participating in the web-based BRAW program during the COVID-19 pandemic. Based on the framework analysis, participants alluded to the importance of the various web-based elements that supported their internalization of the resilience processes. Particularly, the asynchronous and self-paced nature and applicable materials supported participants’ continued engagement with the BRAW program. Finally, after attending the BRAW program, participants became resilient, had greater intent to stay, and were future-ready.

With regard to the web-based elements, the availability of different web-based learning tools has supported participants’ learning. This finding was consistent with prior research that evaluated web-based resilience programs [22,40]. Several key characteristics of web-based learning stood out. First, participants alluded to the importance of short attention-requiring materials such as videos, which was similarly reported in other studies [40,41]. Second, participants credited the availability of quizzes and homework that supplemented their learning. Homework and quizzes can augment the learning process by allowing individuals to apply their newly acquired knowledge [42,43]. Despite the benefits, several participants were overwhelmed by the number of tasks (eg, videos, quizzes, homework, and forum). A unique finding from this study was regarding the nomenclature of the tasks. Particularly, participants mentioned that the term “homework” can be considered a chore and may not be preferred in this form of program. This could be due to participants’ experiences with homework during their schooling years, where numerous negative emotions were associated with that term [44,45].

With regard to the contents, participants credited how the relatability and applicability of the BRAW contents were facilitators for completion. This is an important aspect, as several studies have echoed the importance of providing contextually relevant materials for participants [41,46], and this will facilitate participants’ understanding and transferability of their newly acquired skills. Furthermore, participants appreciated the resilience strategies and applied them in the workplace. For example, the provision of easily replicable strategies such as the application of the problem-solving algorithm was helpful for the participants [27,47].

With regard to the features, the web-based BRAW program was designed as asynchronous and self-paced training for several reasons, such as wider outreach and the presence of the COVID-19 pandemic. The use of a web-based approach was verbalized as an enabling for HCPs to complete the program, which was consistent with other studies [22,48]. In addition, a web-based approach provided HCPs with an opportunity to learn during the COVID-19 pandemic when induced social distancing measures were required. More importantly, the nature of the BRAW program promoted participants’ autonomy and allowed them to gain control over their schedules. This could stimulate personalized learning, which resulted in positive effects on one’s learning outcomes [49,50]. However, despite this, most of the participants also experienced conflicting priorities and were unable to timely participate in the web-based BRAW program. Considering that participation in programs of such nature is of lower priority than their formal work-related commitments, this may have led to their reduced participation [22,27].

Through participants’ narratives, this study also unveiled the positive effects of the web-based BRAW program on their resilience, intent to stay, and employability. From a resilience perspective, the program provided participants with skills ranging from personal (eg, cognitive restructuring), relational (eg, teamwork), and environmental (eg, workplace environment) that promoted their resilience. Based on the resilience theory [11], the introduction of such resilience protective factors can promote resilience. Interestingly, participants’ resilience could also be influenced by the recognition of their resilient potential. Several studies have suggested how the introduction of resilience programs has led to participants becoming aware of their internal strengths and how this influences their resilience [22,51].

Moreover, the web-based BRAW program introduced techniques to enhance cognitive restructuring, positivity, and happiness, and this could be a plausible explanation for improving participants’ intent to stay. Despite the dynamic and stressful health care environment, these techniques potentially supported participants’ positive reframing of a seemingly negative situation [15,31]. Furthermore, it can have positive direct or mediating effects on one’s intent to stay by improving one’s optimism and positivity [52,53]. However, participants also surfaced that macro-organization factors such as hospital administration are factors that may negatively affect their intent to stay [54,55]. While not directly explored in other qualitative evaluations of resilience programs, this study found that the web-based BRAW program has encouraging effects on participants’ employability and future readiness. This could be attributed to the introduction of various noncognitive skills such as problem-solving and
emotion regulation. More literature has highlighted the pivotal role of noncognitive skills on employment outcomes [56,57].

Based on the findings from this qualitative study, several implications for future resilience programs are outlined. First, HCPs continue to experience mental exhaustion and distress due to the immense workload caused by the COVID-19 waves, and the delivery of a web-based program targeting mental well-being is practical and should be implemented. Second, from a feature perspective, an asynchronous and self-paced program is an acceptable and feasible approach. However, to reduce any potential conflicting work commitments, participants should be provided with protected time to complete these programs. Third, web-based learning should be supplemented by various engagement tools, and it will be helpful to redesignate homework as self-help exercises or tasks to reduce the negative connotation associated with homework. Next, from a content perspective, contextualized personal, relational, and environmental resilience materials should be introduced. Thus, conducting a needs analysis would be necessary to ensure that the resilience program remains acceptable to the target population. In addition, there should be an introduction of technical terms for participants who may not be familiar with the materials. Finally, as resilience programs focus on building an individual’s strengths, it will be important that health care administrators consider building supportive workplace environments to complement resilience programs.

Limitations
This study has several limitations, and results need to be interpreted with caution. First, this qualitative study explored participants’ experiences of 1 web-based resilience program, and its findings may not be transferable to other settings. Despite this, our findings may provide insight on the design of future psychosocial web-based interventions. Second, most of them were female and ethnic Chinese participants, thereby resulting in an underrepresentation of other sex and ethnic groups. Nevertheless, a rigorous purposive sampling approach was undertaken to ensure that there is a good representation of individuals across various age groups and professions. Finally, this study was limited to a 1-time point and may not be able to encapsulate the long-term effects of the BRAW program on the participants.

Conclusions
This study presented a qualitative evaluation of a web-based BRAW program using framework analysis. Although there were several highlighted facilitators and barriers, the findings show that an asynchronous, self-paced resilience program can be a useful tool in supporting the well-being of HCPs during the COVID-19 pandemic. However, it will be important to ensure that contextually relevant materials, supported by other appropriate web-based engagement tools, such as quizzes and practical exercises are provided to promote learning in a web-based environment. Further work is needed to explore how macro-organization factors can be embedded in resilience programs to promote HCPs’ resilience and well-being.

Conflicts of Interest
None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File, 25 KB - mededu_v10i1e49551_app1.docx ]

Multimedia Appendix 2
Table S1. Coding tree.
[DOCX File, 18 KB - mededu_v10i1e49551_app2.docx ]

References


44. Agarwal B, Brooks SK, Greenberg N. The role of peer support in managing occupational stress: a qualitative study of the sustaining resilience at work intervention. Workplace Health Saf 2020;68(2):57-64 [FREE Full text]


Abbreviations

**BRAW:** Building Resilience At Work  
**COREQ:** Consolidated Criteria for Reporting Qualitative Research  
**HCP:** health care professional

©Wei How Darryl Ang, Zhi Qi Grace Lim, Siew Tiang Lau, Jie Dong, Ying Lau. Originally published in JMIR Medical Education (https://mededu.jmir.org), 31.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Occupational Therapy Students’ Evidence-Based Practice Skills as Reported in a Mobile App: Cross-Sectional Study

Susanne G Johnson1*, MSc; Birgitte Espenhaug1*, Prof Dr; Lillebeth Larun2*, PhD; Donna Ciliska3*, Prof Dr; Nina Rydland Olsen1*, PhD

1Department of Health and Functioning, Western Norway University of Applied Sciences, Bergen, Norway
2Division of Health Services, Norwegian Institute of Public Health, Oslo, Norway
3Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
*all authors contributed equally

Corresponding Author:
Susanne G Johnson, MSc
Department of Health and Functioning
Western Norway University of Applied Sciences
Inndalaseveien 28
Bergen, 5063
Norway
Phone: 47 92213202
Email: susanne.grodem.johnson@hvl.no

Abstract

Background: Evidence-based practice (EBP) is an important aspect of the health care education curriculum. EBP involves following the 5 EBP steps: ask, assess, appraise, apply, and audit. These 5 steps reflect the suggested core competencies covered in teaching and learning programs to support future health care professionals applying EBP. When implementing EBP teaching, assessing outcomes by documenting the student’s performance and skills is relevant. This can be done using mobile devices.

Objective: The aim of this study was to assess occupational therapy students’ EBP skills as reported in a mobile app.

Methods: We applied a cross-sectional design. Descriptive statistics were used to present frequencies, percentages, means, and ranges of data regarding EBP skills found in the EBPsteps app. Associations between students’ ability to formulate the Population, Intervention, Comparison, and Outcome/Population, Interest, and Context (PICO/PICo) elements and identifying relevant research evidence were analyzed with the chi-square test.

Results: Of 4 cohorts with 150 students, 119 (79.3%) students used the app and produced 240 critically appraised topics (CATs) in the app. The EBP steps “ask,” “assess,” and “appraise” were often correctly performed. The clinical question was formulated correctly in 53.3% (128/240) of the CATs, and students identified research evidence in 81.2% (195/240) of the CATs. Critical appraisal checklists were used in 81.2% (195/240) of the CATs, and most of these checklists were assessed as relevant for the type of research evidence identified (165/195, 84.6%). The least frequently correctly reported steps were “apply” and “audit.” In 39.6% (95/240) of the CATs, it was reported that research evidence was applied. Only 61% (58/95) of these CATs described how the research was applied to clinical practice. Evaluation of practice changes was reported in 38.8% (93/240) of the CATs. However, details about practice changes were lacking in all these CATs. A positive association was found between correctly reporting the “population” and “interventions/interest” elements of the PICO/PICO and identifying relevant research evidence (P<0.001).

Conclusions: We assessed the students’ EBP skills based on how they documented following the EBP steps in the EBPsteps app, and our results showed variations in how well the students mastered the steps. “Apply” and “audit” were the most difficult EBP steps for the students to perform, and this finding has implications and gives directions for further development of the app and educational instruction in EBP. The EBPsteps app is a new and relevant app for students to learn and practice EBP, and it can be used to assess students’ EBP skills objectively.

(JMIR Med Educ 2024;10:e48507) doi:10.2196/48507

KEYWORDS

active learning strategies; application; cross-sectional study; development; education; higher education; interactive; mobile application; mobile app; occupational therapy students; occupational therapy; students; usability; use
Evidence-based practice (EBP) involves using the best available evidence from relevant research and integrating it with clinical expertise, patient values, and circumstances to make clinical decisions for individual patients [1]. When applying EBP, it is recommended to follow the five EBP steps: (1) identifying information needs and formulating answerable questions (ask), (2) finding the best available evidence to answer clinical questions (assess), (3) critically appraising the evidence (appraise), (4) applying the results in clinical practice (apply), and (5) evaluating performance (audit) [1,2]. These 5 steps reflect the suggested core competencies covered in teaching and learning programs to support future health care professionals applying EBP, including developing EBP knowledge and skills [3].

EBP skills can be understood as applying EBP knowledge by performing EBP steps, ideally in a clinical setting [4]. The literature indicates that EBP knowledge and skills improve when EBP teaching and learning are multifaceted, interactive, clinically integrated, and incorporate assessment [5]. When implementing EBP teaching, it is relevant to document and assess the individual student’s performance [3,5,6]. As it is recommended to follow all 5 EBP steps when teaching and learning EBP [1,2], measuring the performance of all 5 steps is relevant when evaluating EBP learning. However, few evaluation instruments measure all 5 EBP steps [5-9], and most instruments are self-reported questionnaires [6,7]. The use of self-reported questionnaires may contribute to biased results due to recall bias or social desirability responses [9,10]. Objectively measuring EBP learning could result in a true reflection of the situation, and thus, it is recommended to develop objective tools for EBP learning assessment [6,7,11]. To objectively document the performance of the EBP steps, Shanefelt et al [6] emphasized using online documentation. Online documentation is feasible through mobile apps, and innovative new methods to evaluate EBP teaching can now be explored [12]. Most students own a smartphone, which makes mobile learning and information sharing possible [13,14]. Thus, mobile apps can potentially be used for documenting and assessing students’ EBP performance. The aim of this study was to assess occupational therapy (OT) students’ EBP skills as reported in a mobile app.

Methods

Design

This study used a cross-sectional design. The reporting of this study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist (Multimedia Appendix 1) [15].

Mobile App

A mobile web app called the EBPsteps app was developed at the Western Norway University of Applied Sciences (HVL) to support health and social care students’ EBP learning [16]. An updated version of this web app is now freely available as a native app [17]. Experiences with using the EBPsteps app for learning EBP have previously been explored [16]. The app provides an opportunity for students to document the 5 EBP steps. A description of the content of the EBPsteps app is presented in Textbox 1.

Textbox 1. The EBPsteps app content.

<table>
<thead>
<tr>
<th>Ask</th>
<th>Reflect on information needs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Formulate the clinical question</td>
</tr>
<tr>
<td></td>
<td>Identify the type of clinical question (drop-down menu)</td>
</tr>
<tr>
<td></td>
<td>Identify the Population, Intervention, Comparison, and Outcome/Population, Interest, and Context (PICO/PICo) elements</td>
</tr>
<tr>
<td>Assess</td>
<td>Report information source used to identify research evidence</td>
</tr>
<tr>
<td></td>
<td>Report links to research evidence identified</td>
</tr>
<tr>
<td>Appraise</td>
<td>Choose a relevant critical appraisal checklist</td>
</tr>
<tr>
<td></td>
<td>Complete the critical appraisal using the integrated checklist</td>
</tr>
<tr>
<td>Apply</td>
<td>Report how research evidence was applied in practice (drop-down menu)</td>
</tr>
<tr>
<td>Audit</td>
<td>Report if changes in practice were completed and evaluated</td>
</tr>
<tr>
<td></td>
<td>Describe changes if changes were implemented</td>
</tr>
<tr>
<td></td>
<td>Evaluate the EBP process (ask, assess, appraise, apply, and audit)</td>
</tr>
</tbody>
</table>
By documenting the EBP process in the app, students produced critically appraised topics (CATs). A CAT can be explained as a summary of research evidence on a clinical question [18]. The CATs completed in the EBPsteps app included information on all EBP steps, and the CATs could be sent through email and shared as a PDF document. The CATs produced in the app were stored on the HVL research server and were accessible to the researchers in this project.

**Participants**

A total of 4 cohorts of fifth-semester OT students from different academic years (from 2018 to 2021) at HVL were eligible for inclusion if they used the EBPsteps app.

**Setting**

In Norway, OT education is a 3-year bachelor’s degree of 6 semesters (180 European Credit Transfer System [ECTS]). According to the Norwegian national curriculum, all health and social care students must be able to acquire new knowledge and make professional assessments, decisions, and actions in line with EBP [19]. At the time of this study, EBP was well integrated into the OT bachelor’s degree program at HVL [20].

Textbox 2 provides an overview of the total number of standalone EBP sessions (n=27) that OT students in this study received by their fifth semester (year 3). This amount of EBP teaching hours is a high number [21]. In addition, EBP was integrated into other learning activities, such as problem-based learning (PBL) group activities, written assignments, and exams.

Using the EBPsteps app was part of the EBP teaching. Students were introduced to the app at the start of the fifth semester. The students watched a video presentation of how to use the app and explored using the app while being supervised by a teacher. During the fifth semester, the students were encouraged to use the EBPsteps app on campus (4 weeks) and during clinical placements (11 weeks). While on campus, students had to use either the EBPsteps app or a Microsoft Word document to complete a mandatory EBP assignment that involved producing a CAT on a clinical topic. Similarly, at the end of the semester, an appendix to the home exam was to use either the EBPsteps app or a Word document to produce a CAT.

**Textbox 2. Overview of standalone EBP sessions. Year 3 includes sessions given through the fifth semester only. EBP: evidence-based practice.**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Standalone sessions about “ask” (2 hours) and “assess” (2 hours). Total duration is 4 hours.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 2</td>
<td>Standalone sessions about “ask” (1 hour), “assess” (1 hour), “appraise” (3 hours), and “apply” (2 hours). Total duration is 7 hours.</td>
</tr>
<tr>
<td>Year 3</td>
<td>Standalone sessions about “ask” (2 hours), “assess” (2 hours), “appraise” (8 hours), “apply” (3 hours), and “audit” (1 hour). Total duration is 16 hours.</td>
</tr>
</tbody>
</table>

**Data Collection**

CATs produced by students during the fifth semester were exported from students’ user accounts in the EBPsteps app to Microsoft Excel [22] at the end of the semester. The Norwegian data, anonymized by authors, are freely available through HVL Open [23] and include our assessment. To objectively assess students’ EBP skills based on how they documented the EBP process in the app, we developed a scoring plan for each EBP step in the CATs (Multimedia Appendix 2). The different steps of the CATs were assessed as correct or incorrect, which were the outcomes investigated in this study. Two researchers independently scored each CAT, and disagreements were resolved through discussion. An overview of the scoring plan is presented in Textbox 3.
Textbox 3. Overview of the scoring plan. Includes the EBP steps and what was assessed. EBP: evidence-based practice.

**Ask**
- Was it reflected on the information needs?
- Which clinical question was formulated (e.g., prevalence, cause, diagnostics, effect of measures, prognosis, or experiences and attitudes)?
- Which clinical question was identified (drop-down menu)?
- Was there an agreement between the formulated clinical question and the type of question identified from the drop-down menu?
- Was the “population” of the Population, Intervention, Comparison, and Outcome/Population, Interest, and Context (PICO/PICo) correctly reported?
- Was the “intervention/interest” of the PICO/PICo correctly reported?
- Was the “comparison” of the PICO/PICo correctly reported?
- Was the “outcome/context” of the PICO/PICo correctly reported?

**Assess**
- Which information sources were used (BMJ Best Practice, Cochrane Library, PubMed, etc)?
- Was a link to research evidence reported?
- Was there an agreement between the information source used and the identified research evidence?

**Appraise**
- Was there an agreement between the identified research evidence and the chosen critical appraisal checklist used?
- Were the questions in the checklist completed?

**Apply**
- Was the application of the research evidence reported (drop-down menu)?
- If reported applied, was this described?

**Audit**
- Were changes in practice evaluated?
- Was the EBP process evaluated?

### Analysis

Descriptive statistics were used to summarize the assessment of students’ EBP skills based on the completed CATs, including frequencies and percentages for categorical variables and mean and range for continuous variables. Associations between correctly reporting the Population, Intervention, Comparison, and Outcome/Population, Interest, and Context (PICO/PICo) elements and finding research evidence were analyzed with the chi-square test with adjustment for repeated measurements [24]. The significance level was set at 5%. Statistical analyses were performed with SPSS Statistics (version 28.0; IBM Corp) [25] and R (R Foundation for Statistical Computing) [26].

### Ethical Considerations

The Norwegian Agency for Shared Services in Education and Research approved the study (project 50425). The students were informed, both orally and in writing, about the purpose of this study and that the data would be treated confidentially. The students agreed to participate in the study and signed a consent form when they created a profile and used the EBPsteps app. The students did not receive any compensation for participating. Students could choose to use the app or a Word document to complete assignments where it was required to produce CATs. The data were securely stored on the research server at HVL.

### Results

#### Participants

Among 4 cohorts with OT students, 79.3% (119/150) of students used the EBPsteps app during their fifth semester. The students who used the app produced 240 CATs. In the first cohort (2018), 41 of 47 students produced 73 CATs; in the second cohort (2019), 25 of 30 students produced 53 CATs; in the third cohort (2020), 21 of 33 students produced 43 CATs; and in the fourth cohort (2021), 32 of 40 students produced 71 CATs. The mean number of CATs produced per student was 2, with a range from 1 to 7.

#### Step 1: Ask

A need for more knowledge on a clinical problem was reported in 94.6% (227/240) CATs. In 80% (192/240) of the CATs, the type of clinical question was identified using a drop-down menu. A clinical question was formulated in 53.3% (128/240) of the CATs. The "effect of therapy" was the most prevalent clinical question reported (100/240, 41.7%) (Table 1).
All PICO/PICo elements were reported correctly in 10.4% (25/240) of the CATs. Assessing the different PICO/PICo elements separately, the “population” and “intervention/interest” elements were more often correctly reported (187/240, 77.9% and 189/240, 78.8%) than the “comparison” and “outcome/context” elements (44/240, 18.3% and 103/240, 42.9%). This applied to all question types, including when the question had been formulated as a background question (Table 1). In CATs without a clinical question identified, most PICO/PICo elements were incorrectly reported.

**Table 1.** Correctly reported Population, Intervention, Comparison, and Outcome/Population, Interest, and Context (PICO/PICo) elements by type of question in 240 critically appraised topics.

<table>
<thead>
<tr>
<th>Effect of therapy (n=100)</th>
<th>Qualitative (n=27)</th>
<th>Background (n=64)</th>
<th>Other (n=1) or missing (n=48)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population, n (%)</td>
<td>Intervention/interest, n (%)</td>
<td>Comparison, n (%)</td>
<td>Outcome/context, n (%)</td>
</tr>
<tr>
<td>90 (90)</td>
<td>96 (96)</td>
<td>30 (30)</td>
<td>53 (53)</td>
</tr>
<tr>
<td>25 (93)</td>
<td>25 (93)</td>
<td>N/R⁴</td>
<td>13 (48)</td>
</tr>
<tr>
<td>55 (86)</td>
<td>52 (81)</td>
<td>11 (17)</td>
<td>32 (50)</td>
</tr>
<tr>
<td>17 (35)</td>
<td>16 (33)</td>
<td>3 (6)</td>
<td>5 (10)</td>
</tr>
</tbody>
</table>

⁴Not relevant.

**Step 2: Assess**

In 240 of the CATs, the information source most frequently reported was the Cochrane Library (65/240, 27.1%), followed by CINAHL (43/240, 17.9%), PubMed (36/240, 15%), and Epistemonikos (17/240, 7.1%). In 12.9% (31/240) of the CATs, no information source was reported. Research evidence was identified and linked to in 81.3% (195/240) of the CATs, and the most common type of research evidence identified was systematic reviews (n=85), randomized controlled trials (RCTs; n=51), and qualitative research (n=44).

We observed a positive association between correctly reporting “population” and “intervention/interest” elements of the PICO/PICo and identifying research evidence. Among those correctly reporting the population element, 92.1% (221/240) identified research evidence, compared to 52.1% (125/240) among those that did not report the population element (P<.001). Similar findings were observed for the intervention/interest element.

**Step 3: Appraise**

A checklist was used in 81.3% (195/240) of the CATs. Of these, the correct checklist was used in 84.6% (165/195) of the CATs; that is, there was agreement between the type of checklist and the research evidence identified (Table 2). In 98.2% (162/165) of the CATs with a correct checklist, more than 75% of the checklist questions had been answered. Effect estimates from identified research evidence were documented in 27% (21/77) of the checklists for systematic reviews and 36% (15/42) of the checklists for RCTs.

**Table 2.** Type of research evidence identified and agreement with choice of checklist.

<table>
<thead>
<tr>
<th>Type of research evidence</th>
<th>The agreement between research evidence and checklist, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic reviews (n=85)</td>
<td>77 (89)</td>
</tr>
<tr>
<td>Randomized controlled trials (n=51)</td>
<td>42 (82)</td>
</tr>
<tr>
<td>Qualitative research (n=44)</td>
<td>42 (95)</td>
</tr>
<tr>
<td>Guidelines (n=4)</td>
<td>2 (50)</td>
</tr>
<tr>
<td>Observational studies⁵ (n=11)</td>
<td>2 (18)</td>
</tr>
<tr>
<td>The total number of research evidence identified (n=195)</td>
<td>165 (84.6)</td>
</tr>
</tbody>
</table>

⁵Included the following study designs: prevalence (n=1), diagnostic (n=1), cohort (n=3), case-control (n=1), and cross-section (n=5).

**Step 4: Apply**

In 39.6% (95/240) of the CATs, it was reported that research evidence was applied in clinical practice. How the research was applied was described sufficiently in only 61% (58/95) of these CATs.

The most common shared decision-making approach reported from a drop-down menu was “identifying preferences” (78/240, 32.5%) and “exploring possibilities” (78/240, 32.5%). Other shared decision-making approaches reported were “presenting choices” (48/240, 20%) and “recommendations” (46/240, 19.2 %), “discussing potential” (45/240, 18.8%), “deciding follow-up” (28/240, 11.7%), and “checking recommendations” (24/240, 10%).

**Step 5: Audit**

Evaluation of practice changes was reported in 38.6% (93/240) of the CATs. However, details of practice changes were lacking in all these CATs. In 46% (43/93) of the CATs that reported evaluation, it was reported, “did not change practice,” and in 54% (50/93) of these CATs, it was reported that it was “not relevant to change practice.” The EBP process was reported as evaluated in 54.6% (131/240) of the CATs.
**Discussion**

**Principal Findings**
This study assessed OT students’ EBP skills as reported in the EBPsteps mobile app. We found that students were most often able to perform the EBP steps of “ask,” “assess,” and “appraise” correctly. A positive association was found between formulating the PICO/PICo elements and identifying research evidence. Applying the evidence and evaluating practice change were the least frequently correctly reported steps of the EBP process.

**Comparison to Previous Work**
Using data from the EBPsteps app, where students had documented how they followed the EBP process for their clinical question, enabled us to collect objective data on students’ EBP skills. Instruments that objectively measure EBP skills are recommended for acquiring a true reflection of the situation [6,7,11], as opposed to more frequently used self-report assessment tools [6,7]. Although objective assessment is advised, it can be time-consuming to complete and assess [4]. Consequently, self-reported questionnaires are often chosen because of their practicality of administration [9]. Developing an easy-to-administer scoring plan for the EBPsteps app has therefore been important. Against this background, the EBPsteps app can be a valuable contribution to objectively assessing EBP skills related to all 5 steps of the EBP process.

**Ask and Assess**
We found a positive association between correctly reporting population and intervention/interest elements of the PICO/PICo and finding research evidence, indicating that completing the PICO/PICo supports students’ ability to retrieve relevant research evidence. These findings align with previous research reporting that a clearly defined question supports students’ ability to retrieve relevant information [27,28]. Furthermore, structuring the question using the PICO/PICo format makes it easier to decide on search terms [2].

**Appraise**
The appropriate critical appraisal checklist was chosen in 68.8% (165/240) of the CATs in this study. Nevertheless, few effect estimates were reported in checklists for RCTs and systematic reviews. This might suggest that the students had difficulties interpreting the statistical results. Lack of confidence in interpreting statistical results has previously been reported among health and social care students [29,30]. Acquiring an understanding of effect estimates is necessary when applying EBP [3], and spending more time teaching the understanding of research results to support the students’ learning and interpretation of research results is recommended [31].

**Apply and Audit**
Only about half of the students in this study reported that they applied the research evidence they found, indicating that they struggled using EBP skills beyond the classroom setting, which also correlates with previous research [32,33]. Lehane et al [34] suggest that structural incorporation of EBP during clinical placement, for instance, through easy access to research, EBP mentors, or regular journal clubs, may support the students in applying research evidence. In addition, incorporating assessment of EBP into clinical placement has been shown to influence EBP behavior [5]. In this study, EBP assignments were mandatory in class but not during clinical placement, which may explain why students in this study struggled with the steps of applying and evaluating practice. Providing a mandatory EBP assignment during the clinical placement may support the students in applying EBP and thus also mastering the 2 last steps of the EBP process.

An alternative explanation for why students struggled with the steps of applying and evaluating practice could be that they experienced fatigue or other difficulties using the app. To explore whether other issues influenced students’ skills, we could have further tested the usability of the app. When developing mobile apps for teaching and learning, usability testing is important [35]. Other research methods are necessary to investigate why the 2 last steps of the EBP process were less frequently completed. Future research should include cognitive interview studies (eg, think-aloud methods) and other pilot studies in different populations to evaluate the comprehensiveness and comprehensibility of the app.

**Future Directions**
Knowledge of which EBP steps students find most challenging has implications and gives directions for further development of the EBPsteps app and educational instruction in EBP. For example, providing a more comprehensive explanation of how to interpret statistical results in the app could be beneficial. In addition, spending more time teaching statistics and how to read the results seems necessary to improve students’ EBP performance.

A better alignment between what is taught during classes on campus and what students do at placements could also perhaps better facilitate EBP behavior among students. A mandatory assignment where research evidence must be found and discussed with the clinical instructors may help the students apply and evaluate the use of research evidence during clinical placement.

Currently, the EBPsteps app is available only in Norwegian. In the future, we aim to provide user interface translations for several languages [16]. However, we will need to modify options in the app according to the free access resources available in the different countries (eg, databases, guidelines, and e-learning resources). Efforts will be made to find the best solution and to accommodate needs in low- and middle-income countries.

**Methodological Considerations**
The main limitation of this study was that we included students from only one profession and from the same educational institution, and thus the generalizability of the results to other institutions and to other health and social care students is reduced. However, the sample consisted of 4 student cohorts from different academic years (from 2018 to 2021; n=119), including 240 CATs. Accordingly, we believe the results from this study can be recognizable and relevant across other populations.
A strength of this study was that the EBPsteps app allowed us to objectively measure the performance of the EBP process using an app that includes all 5 EBP steps. It is recommended that educators select instruments that objectively measure EBP performance [11]. Shaneyfelt et al [6] emphasized the use of online documentation of the EBP steps as a promising approach.

Another strength was that 2 researchers assessed the CATs independently based on a scoring plan, and disagreement was solved through discussion. However, the EBPsteps app and the scoring plan are not validated for assessing EBP, and measurement properties should be examined in future studies.

Conclusions
We assessed the students’ EBP skills based on how they documented following the EBP steps in the EBPsteps app, and our results showed variations in how well the students mastered the steps. “Apply” and “audit” were the most difficult EBP steps for the students to perform, and this finding has implications and gives directions for further development of the app and educational instruction in EBP. The EBPsteps app is a new and relevant app for students to learn EBP and can be valuable for assessing EBP skills objectively.

Acknowledgments
The authors would like to thank Johannes Mario Ringheim at Medialab, HVL, for the programming and technical development of the EBPsteps app and data extraction from the EBPsteps app for this study. In addition, the authors would like to thank all the students who participated in the study and used the EBPsteps app.

Data Availability
The Norwegian data, anonymized by the authors, are publicly and freely available through HVL Open [23].

Authors’ Contributions
SGJ and NRO conceptualized this study. NRO was responsible for the funding of the study, and the initial analysis of the results and the project administration were performed by SGJ and NRO. The formal analysis was conducted by SGJ and BE. SGJ, BE, LL, DC, and NRO decided on the methodology. SGJ, BE, and NRO provided resources. Validation was done by SGJ, BE, and NRO, and visualization by SGJ and NRO. The writing of the original draft was done by SGJ, and review and editing were done by SGJ, BE, LL, DC, and NRO.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist.
[DOC File, 113 KB - mededu_v10i1e48507_app1.doc ]

Multimedia Appendix 2
The scoring plan of EBPsteps.
[DOCX File, 38 KB - mededu_v10i1e48507_app2.docx ]

References


17. EBPsteps. Western Norway University of Applied Sciences. 2015. URL: https://www.ebpsteps.no [accessed 2023-08-29]


24. Gregg M, Datta S, Lorenz D. R package version 0.2.2. htestClust: Reweighted Marginal Hypothesis Tests for Clustered Data. URL: https://cran.r-project.org/web/packages/htestClust/index.html [accessed 2023-05-01]


Abbreviations

CAT: critically appraised topic
EBP: evidence-based practice
ECTS: European Credit Transfer System
HVL: Western Norway University of Applied Sciences
OT: occupational therapy
PBL: problem-based learning
PICO/PICO: Population, Intervention, Comparison, and Outcome/Population, Interest, and Context
RCT: randomized controlled trial
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
Evaluating the Impact of the National Health Service Digital Academy on Participants’ Perceptions of Their Identity as Leaders of Digital Health Change: Mixed Methods Study

Amish Acharya1*, MBBS, BSc; Ruth Claire Black1*, EdD, JD; Alisdair Smithies1, PhD; Ara Darzi1, MD, FRCS
Institute of Global Health Innovation, Imperial College London, London, United Kingdom
*these authors contributed equally

Abstract

Background: The key to the digital leveling-up strategy of the National Health Service is the development of a digitally proficient leadership. The National Health Service Digital Academy (NHSDA) Digital Health Leadership program was designed to support emerging digital leaders to acquire the necessary skills to facilitate transformation. This study examined the influence of the program on professional identity formation as a means of creating a more proficient digital health leadership.

Objective: This study aims to examine the impact of the NHSDA program on participants’ perceptions of themselves as digital health leaders.

Methods: We recruited 41 participants from 2 cohorts of the 2-year NHSDA program in this mixed methods study, all of whom had completed it >6 months before the study. The participants were initially invited to complete a web-based scoping questionnaire. This involved both quantitative and qualitative responses to prompts. Frequencies of responses were aggregated, while free-text comments from the questionnaire were analyzed inductively. The content of the 30 highest-scoring dissertations was also reviewed by 2 independent authors. A total of 14 semistructured interviews were then conducted with a subset of the cohort. These focused on individuals’ perceptions of digital leadership and the influence of the course on the attainment of skills. In total, 3 in-depth focus groups were then conducted with participants to examine shared perceptions of professional identity as digital health leaders. The transcripts from the interviews and focus groups were aligned with a previously published examination of leadership as a framework.

Results: Of the 41 participants, 42% (17/41) were in clinical roles, 34% (14/41) were in program delivery or management roles, 20% (8/41) were in data science roles, and 5% (2/41) were in “other” roles. Interviews and focus groups highlighted that the course influenced 8 domains of professional identity: commitment to the profession, critical thinking, goal orientation, mentoring, perception of the profession, socialization, reflection, and self-efficacy. The dissertation of the practice model, in which candidates undertake digital projects within their organizations supported by faculty, largely impacted metacognitive skill acquisition and goal orientation. However, the program also affected participants’ values and direction within the wider digital health community. According to the questionnaire, after graduation, 59% (24/41) of the participants changed roles in search of more prominence within digital leadership, with 46% (11/24) reporting that the course was a strong determinant of this change.

Conclusions: A digital leadership course aimed at providing attendees with the necessary attributes to guide transformation can have a significant impact on professional identity formation. This can create a sense of belonging to a wider health leadership structure and facilitate the attainment of organizational and national digital targets. This effect is diminished by a lack of locoregional support for professional development.

(JMIR Med Educ 2024;10:e46740) doi:10.2196/46740
KEYWORDS
digital leadership; professional identity; dissertation of practice

Introduction

Background

Delivering the digital transformation of the United Kingdom’s National Health Service (NHS) has been a long-standing aim. In the “What Good Looks Like” framework, by 2025, the NHS aims to have all integrated care systems and associated trusts reach core digital capability [1]. The key to this digital leveling-up strategy is the need to support professional development and training opportunities across integrated care systems [2]. To facilitate system-wide progress, there is a growing need for digitally proficient leadership teams; however, one of the main barriers identified by the NHS Transformation Directorate has been a lack of a “clear steer” for digital decisions [3].

Digital health resources and digital tools that were adopted through necessity during the COVID-19 pandemic have led to a paradigm shift in routine care. The scope of these resources has been significant across health care, including remote patient-clinician consultations and diagnostics [4-6]. However, as the health care service looks to enact facets of the NHS Long Term Plan and scale future sustainable digital change, possessing robust leadership to set this direction is key [7].

The NHS Digital Academy (NHSDA) designed its flagship course to deliver this support to emerging leaders. Each cohort of approximately 100 professionals is selected from applicants who are directly employed by the NHS or social care in England [8]. Digital health leadership is delivered in 2 accredited components. The first, resulting in a Postgraduate Diploma (PGDip) in Digital Health Leadership, uses a blended learning approach to provide a theoretical framework for topics such as user-centered design [9,10]. This involves web-based teaching on 6 core modules structured around assessment deadlines including the essentials of health systems, implementing change, health information systems, user-centered design, actionable data analytics, and leadership change. Subsequently, students can undertake a 1-year Master of Science (MSc) degree. The MSc degree uses a dissertation of the practice model, where students focus on practicable applications of theory within defined digital transformation projects. This self-directed period of study involves candidates’ leading projects within their own host organizations with periodic deadlines to guide progress and continued access to the support of the teaching faculty. In this manner, the course facilitates a workplace-based learning model geared toward supporting students to use research to solve a real-world problem in their organization. Although the course has been shown to effectively impact the attainment of national digital priorities [11], little is known about the effect on participants’ perceptions of themselves as digital leaders or their professional identity.

Defining professional identity is difficult owing to a lack of standardization of the term. It has been associated with knowledge acquisition, performance of typical tasks, displays of expected behaviors, or shared ethos and value systems [12]. In other contexts, professional identity involves the integration of the personal and professional selves [13]. A scoping review by Cornett et al [14] identified 5 constructs associated with health professional identity, including lived experience (eg, practicing), the world around me (eg, the workplace), belonging (eg, collective identity), me (eg, self in relation to the profession), and learning (eg, acquiring skills). However, the review examined health professional practice and did not specifically examine digital leaders [14]. Understanding what constitutes the identity of a digital health leader is potentially more problematic, given that the field is relatively nascent. Consequently, understanding what knowledge is needed, the tasks or behaviors that are expected of leaders, and what constitutes core values is likely to remain ill-defined until the digital health landscape has evolved. Furthermore, an individual’s perception of their professional role is a dynamic process and can be augmented by one’s context [15]. Determinations regarding the extent to which one feels like a professional can therefore be difficult to ascertain.

Despite these challenges, professional identity formation has been shown to be an increasingly important aspect of learning development. Within clinical settings, professional identity contributes to the delineation of practice boundaries as well as avoiding confusion regarding individuals’ roles within wider teams [16]. With a growing body of clinicians involved in digital leadership, this is particularly important, as studies have demonstrated that doctors can often encounter difficulties when reconciling managerial and clinical responsibilities. Moreover, aspects of professional identity, such as “belongingness,” have been associated with greater workforce retention [17]. Therefore, there is a growing drive to evaluate how courses and educational curricula impact an individual’s identity.

Aim

This mixed methods study aimed to understand the influence of the NHSDA Digital Health Leadership program on participants’ perceptions of themselves as digital health leaders. This will facilitate a greater understanding of the core values associated with digital leadership and provide insights to improve courses globally.

Methods

This study was conducted as a mixed methods study involving a web-based questionnaire, interviews, and focus groups.

Recruitment

Participants in the first 2 cohorts of the NHSDA’s flagship Digital Health Leadership program were recruited for the study. All participants had completed both years of the program and were >6 months from completion to avoid recency bias. This could involve overemphasizing the impact of later teaching in course compared to that which occurred earlier. Studies suggest that a later evaluation can provide a more holistic evaluation [18]. It also provided time for candidates to reflect on future career opportunities. No other exclusion criteria were placed...
upon participants; therefore, a nonprobabilistic sampling method to reach the necessary sample size was used. Eligible participants were contacted through email by a member of the research team (AA) with no direct link to the NHSDA. Both cohorts were impacted by the COVID-19 pandemic, particularly with respect to their dissertation projects that were undertaken during the pandemic.

**Scoping Questionnaire**

A previously validated web-based scoping questionnaire was used to provide insights and feedback on the course [11]. This questionnaire explored the impact of the course on the development of facets such as "social intelligence," "interpersonal skills," and "courage." It also examined the effect of the course on future goals, asking "Would you consider any of the following additional training options in Digital Health Leadership or a related field within the next 2 years?" This questionnaire was developed to map specifically onto the NHSDA program objectives and encompassed questions including individuals’ perspectives on development and digital leadership. It also sought to ascertain feedback on the aspects of the course that were most influential on participants. A total of 2 authors (RCB and AA) developed the survey questions, whereas a third (AS) independent author was involved to discuss disagreements. The participants were recruited via an email containing an anonymous link. The links were delivered to all eligible individuals separately from the program to avoid selection and response biases based on prior performance in the course.

**Semistructured Interviews**

Following the survey, anonymous responses were quantitatively (multiple-choice questions) and qualitatively analyzed (free-text sections). Themes derived from the analysis were elicited by 2 authors (RCB and AS). The results from the survey were then used to develop the question guides for interviews and focus groups (Multimedia Appendix 1). A third author (AS) was involved to help resolve disagreements. Interviews were designed to gain a more in-depth understanding of individuals’ perceptions of the values and skills associated with digital health leadership and how the course has influenced these areas. Enrollment to an interview was not dependent on completion of the survey or prior performance. All interviews were conducted web-based via Microsoft Teams (Microsoft Corporation). AA conducted all the semistructured interviews, and AA had no formal role within the NHSDA and no prior interaction with any participants to avoid response biases.

**Focus Groups**

To ascertain the shared experience of participants and paralleling the collaborative learning approach used by the program, web-based focus groups were also undertaken using the Microsoft Teams platform. In addition, the focus groups examined the participants’ contrasting experiences of the course. A total of 3 focus groups were conducted, with the facilitator (AA) not being affiliated with the NHSDA. Each focus group involved 4 to 5 participants. The invitation to participate in the focus groups was not contingent on the completion of any previous phase of the study. As with the interviews, the focus groups used open-ended prompts to foster responses. In addition, the facilitator encouraged open discussion between participants. Identity involves the development of attributes congruent with the profession, that is, a common set of values about what it means to be a digital health leader [19]. By facilitating focus group discussions regarding how digital leadership is perceived, its underpinning principles, and how one can develop the necessary skills to become a more effective leader, a greater understanding of these shared values was attained.

**Analysis**

Survey responses were collated through the web-based tool Qualtrics (Qualtrics International Inc). Qualtrics automatically aggregates replies and provides frequencies from the respondents by choice. Given the small number of responses and because the initial survey was used to inform further study phases, no statistical analysis was undertaken. Free-text options were inductively thematically analyzed until data saturation was achieved by an author (AA) and validated by another (RCB). Both qualitative and quantitative responses were used to inform the development of the topic guides following discussion between the authors. Specifically, the authors focused on areas of disagreement or if a particular topic recurred across the responses of different participants.

Audio recordings, obtained with the consent of participants for both interviews and focus groups, were transcribed using the web application Descript (Descript, Inc). The accuracy of the outputs was confirmed by one author (AA), who was present in the interviews and focus groups. Anonymous transcripts were then uploaded to the analysis tool MAXQDA (VERBI GmbH). A deductive thematic analysis was conducted using a technique previously used in similar studies [20]. This involved familiarization with the transcripts by 2 authors (RCB and AA). The transcripts were then coded with the data explored to examine the frequency and relationship of the codes. Similar codes were combined into themes and subthemes, which were aligned with the components of professional identity elicited by Chin et al [19]. This review was selected as a framework on which to base the thematic analysis for 3 reasons: first, because of its comprehensive evaluation of identity with 10 evidence-based facets described; second, the examination of internship or workplace-based learning parallels the educational model of the NHSDA's second year; and finally, the authors’ examination of how these components map to other contexts can facilitate cross-discipline comparisons was helpful in understanding the participant’s identity across wider teams [21]. Although Chin et al [19] found that only a subset of these components was applicable to higher education internships, this study examined the relevance of all components, as some were more significant in postgraduate studies. As a means of validation, the anonymized transcripts were reviewed again, and themes were amended until a consensus was attained. All discrepancies within the coding exercise or allocations of themes were discussed until resolution. Themes that were consistently mentioned by different participants, those that aligned with findings from the questionnaire or focus group, and those that were regarded as stronger determinants were considered more impactful influences. A constructivist approach
was used as the basis of this study, which paralleled the active learning undertaken throughout the program. The paradigm focuses on the importance of active learning and its transformation through experience [22,23]. It involves the engagement and reflection of the learner, which can be impacted by context, knowledge, motivation, values, or organizational setting [24]. This is particularly pertinent to identity development, which can be influenced by such intrinsic and extrinsic factors.

High-scoring dissertations across the 2 included cohorts were also evaluated independently by 2 authors (RCB and AA). The authors then mapped the skills exhibited in these manuscripts to the components of professional identity. Students were required to make explicit reference to a particular component for it to be mapped. The authors discussed any disagreements until a consensus was reached.

**Ethical Considerations**

Approval to conduct this study was provided by the Institutional Review Board at Imperial College London (reference EERP2021-026a). All participants provided explicit written consent to participate in the study and were free to withdraw at any time. No participant received financial remuneration for being involved in the study. All data including transcripts and survey data were kept anonymous, in keeping with the secure data storage policies of Imperial College London.

**Results**

**Overview**

A total of 41 eligible participants completed the web-based survey, of which 42% (17/41) were female and 59% (24/41) were male. Most participants were in clinical health care roles (17/41, 42%), whereas 34% (14/41) were in program delivery or management roles; 20% (8/41) were in informatics or data science roles, and 5% (2/41) were in “other” roles. Of those surveyed, 59% (24/41) reported that the NHS Digital Academy course had a strong and direct impact on their working practice, 27% (11/41) reported some impact, and only 2% (1/41) reported no effect. In total, 4 key themes were elicited from the inductive analysis of the free-text sections: transformative impact, valuing collaboration, goal setting, and improving positive perceptions. The selected results are presented in Multimedia Appendix 1.

Semistructured interviews were conducted with 34% (14/41) of participants. The demographics of which paralleled those from the wider cohort, with 43% (3/7) of participants identifying as female and 50% (7/14) working in clinical roles. In total, 3 focus groups were held with more than half of the attendees (7/13, 54%) not involved in the preceding interviews. The data sources including the number of participants used in the study is presented in Figure 1.
Thematic analysis mapped findings from interviews and focus groups to 8 of the 10 components of professional identity highlighted by Chin et al [19]. Internship experience was not measured, as most of the cohort had been in their roles before the NHSDA and could not be considered entering an internship. However, aspects encompassing skill acquisition during dissertations were covered in other domains. The work environment was also not included, as the participants came from disparate fields, precluding comparisons. However, the findings were mapped to the following domains: commitment to the profession, critical thinking, goal orientation, mentoring, perception of the profession, socialization, reflection, and self-efficacy. When undertaking the thematic analysis and mapping of the highest-scoring dissertation to the components from the framework by Chin et al [19], only 4 were found to be applicable. These included critical thinking, goal orientation, mentoring, and reflection. This is likely because the dissertation was more descriptive of a specific transformation project rather than reflective of the attitudes of participants toward digital health leadership as a whole. 

Table 1 demonstrates how the course impacted these areas through quotations from the respondents.
eldicated this transformative impact of the course upon decision. Inductive thematic analysis of the survey comment participants who changed their roles following the course, 46% reported changing their roles following the course. Among the program appeared to influence individuals' commitment to the profession

<table>
<thead>
<tr>
<th>Domain</th>
<th>Definition</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commitment to profession</td>
<td>The physical, mental, and emotional commitment to being a digital leader.</td>
<td>• “People saw what I was learning from the course and my enthusiasm for my career...I was invigorated and it encouraged me to do the best in my career path.” [Interviewee 13]</td>
</tr>
<tr>
<td></td>
<td>Understanding the aims of leadership and demonstrating a willingness to achieve them.</td>
<td>• “Having gone through the course, I want to pursue careers in digital in some fashion...it’s shifted the course of my career, I’m aspiring to national level roles.” [Interviewee 7]</td>
</tr>
<tr>
<td>Critical thinking</td>
<td>A metacognitive skill to critically evaluate the current standard and elicit new solutions. Involves understanding one’s own role as a digital leader and how one would want it to be.</td>
<td>• “There is a critical language element to it [digital leadership]...you have more critical analysis aspect to your work...it [the National Health Service Digital Academy] has changed the way I approach and think about problems.” [Interviewee 6]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• “I became more strategic in my approach...it’s broader...it’s about thinking right...we need it think from top to bottom.” [Interviewee 5]</td>
</tr>
<tr>
<td>Goal orientation</td>
<td>How one achieves and defines the specific outcomes associated with being a digital leader. Involves having a conducive environment for task mastery and development as a digital leader.</td>
<td>• “My perception has shifted...I see myself as a facilitator of digital transformation...I aim to maintain virtual delivery and my organization is helping me meet that need.” [Interviewee 11]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• “One of the most valuable things from the Digital Academy...it made me understand where to make change, improve processes, how to measure that change and feeding it back...it is the core of our aims.” [Interviewee 1]</td>
</tr>
<tr>
<td>Mentoring</td>
<td>Acting as a mentor and having mentorship. Involves role modeling, feedback strategies, and encouraging self-reflection as a digital leadership, as well as a conducive work environment for mentor.</td>
<td>• “I believe in paying it forward, I’ve brought back what I’ve learnt to building my informatics team.” [Interviewee 7]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• “I became very invigorated by the community...I spent an afternoon with a module lead in user design... then the head of user design centre in the NHS offered me an opportunity to shadow them.” [Interviewee 13]</td>
</tr>
<tr>
<td>Perception of the profession</td>
<td>Ideas about what it is to be a digital leader, the skills required, and its place in wider health care infrastructure.</td>
<td>• “The NHSDA meant I didn’t hold those people on a pedestal...it [digital leadership] is not about having all the technical knowledge, it’s being able to pull together everyone toward the solution.” [Interviewee 11]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• “It has been transformational...just in the knowledge it has given me...on understanding the role...where it fits into organizational strategy...the scope.” [Interviewee 4]</td>
</tr>
<tr>
<td>Professional socialization</td>
<td>A sense of belonging to the wider community and being accepted as part of a group. Includes credentialing and peer networks.</td>
<td>• “It has a level of kudos...there is good recognition that it, the academy trained them up.” [Interviewee 12]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• “Now I’ve got a network of probably 100 or more contacts nationally...I would go and talk to them and say you must know someone locally who does this, any chance you could put me in touch?” [Interviewee 8]</td>
</tr>
<tr>
<td>Reflection</td>
<td>Reflecting on knowledge, cognition, professional identity, maturity, and the sense of professionalism within digital leadership. Involves ideas regarding professional development.</td>
<td>• “It’s highlighted the positives and the negatives of my leadership style, my digital knowledge and also where I fit within an organization and nationally. So it’s given me that sort of self-awareness.” [Interviewee 10]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• “It helped me become better leader. It helped me understand how would I help people in the organization transform and be more innovative.” [Interviewee 2]</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>Self-belief or belief in one’s own capabilities to perform as a digital leader. Includes impostor syndrome and the impact of external opinions upon one’s own beliefs.</td>
<td>• “When I first went on the digital academy...it felt like we were interlopers...throughout it continued to build my confidence levels and where I fit as a digital leader locally.” [Interviewee 14]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• “It made me think that I am a leader...I would never have applied for that Royal College job without it.” [Interviewee 3]</td>
</tr>
</tbody>
</table>

aWork environment not included as relevant components covered in “mentoring,” and participants came from disparate environments.

bInternship experience was not included. The participants represent the existing digital health leadership whose roles would not include internship. Areas of skill-building are covered within other domains.

**Commitment to the Profession**

The program appeared to influence individuals’ commitment to digital health leadership. In total, 59% (24/41) of the cohort reported changing their roles following the course. Among the participants who changed their roles following the course, 46% (11/24) reported that the program had a strong impact on this decision. Inductive thematic analysis of the survey comment elicited this transformative impact of the course upon candidates’ careers, with several describing “life-changing” or “career-changing” effects. In one focus group, one informatician mentioned that “I do now feel like a leader, and I wasn’t going to stay in that organisation.” This new commitment led them to “find somewhere else that I [they] could be a digital leader.” Others reported that they “were looking at influencing policy, in a way I [they] hadn’t before...because of the course.” This commitment to digital health leadership has led them to apply for chief clinical informatics officer (CCIO) roles. The course
also appeared to reaffirm participants’ motivation for undertaking digital health leadership roles. One CCIO stated:

[The course] hasn’t necessarily given me all the technical skills...but it’s greater than that. It’s given the background of how we’ve got to where we are now and inspired me to change things going forward.

Critical Thinking

Critical thinking, which involves understanding a context and deriving new solutions, was found to be fostered predominantly through the MSc dissertation. As presented in Table 2, all but 3 of the 30 highest-ranking dissertation topics across the 2 cohorts involved critical analysis.

Providing a supportive environment for change enabled candidates to put theoretical learning into practice. An interviewee said “I [they] approach things differently, I’m [they are] more strategic, more constructed after the project.” These cognitive skills have continued postgraduation with individuals feeling they have “different tools that were picked up during the academy, which I [they] use day-to-day.”
Table 2. Topics of the highest-scoring dissertations of the 2 cohorts and components of identity that were incorporated.

<table>
<thead>
<tr>
<th>Dissertation topic</th>
<th>Critical thinking</th>
<th>Goal orientation</th>
<th>Mentoring</th>
<th>Reflection</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYOD policy design and development for NHS Trusts</td>
<td>✓✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Board level digital readiness</td>
<td></td>
<td>✓✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blueprint for digital excellence in the development of a new hospital</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementing recommendations of Topol review</td>
<td></td>
<td>✓✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impact of digital working on patient care</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improving performance of a cardiorespiratory outpatient department</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>App to reduce suicide and self-harming and improve safety and clinical outcomes in mental health</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standards and processes for sharing data across platforms and organizations</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blueprint for digital first GP</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant preferences for contact and clinical research study enrollment</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluating impact of digital maturity on effectiveness and efficiency of care in adolescent inpatient mental health units</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital transformation of epilepsy care and monitoring</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementing SNOMED-CT coding into an EHR for clinical decision support, data sharing and medical pathway transformation</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct web-based advice from consultant psychiatrists to GPs</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Returning health professionals living with cancer to work via a digital resource</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enabling effective and appropriate use of virtual consultations with adolescents in psychiatry specialty settings</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An impact analysis of Morse system implementation and mobile device use by health visitors in rural Scotland</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimizing remote access to primary care during COVID-19: a focus on patients with moderate to severe mental health needs</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Making quite voices louder: addressing health inequalities for people with moderate to severe mental health illness</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digitally enabling primary care beyond the COVID-19 pandemic</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The impact of digital tools and ways of working on staff burnout and enjoyment of work in psychiatry</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimizing culture of collaboration and learning to tackle health inequalities: a study of digital health Canada</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital delivery: the future of UK diabetes education</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impact of the implementation of a critical care information system on patient-facing clinical staff in an intensive care unit during the COVID-19 pandemic</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The key components of organizational culture for a digital first strategy</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The relationship between funding and the digital maturity of NHS provider organizations</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partnership between health care provider organizations and industry in adopting AI into health care practice</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A framework for effective prioritization of digital transformation projects in recently merged secondary care organizations</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best practices for digital inclusion in at risk pediatric populations</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformation at pace and scale by EPR sharing among high and low digitally mature hospital systems</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aBYOD: bring your own device.
Goal Orientation

Goal orientation encompasses defining and accomplishing the specific outcomes of digital leadership within the NHS. This may involve achieving national priorities outlined in health policies, such as the NHS Long Term Plan or locoregional transformation targets. As the MSc model was designed to provide a supported environment to undertake these projects, it was unsurprising that all the dissertations evaluated involved an element of goal orientation. These projects varied from digitizing diabetes education platforms and booking processes to projects focusing on the implementation of the recommendations of the Topol review [25]. An interviewee said, “The reason I [they] chose this MSc project was because...it was my day job.” This pragmatic approach helped align the goals of the course with those of digital leaders. Survey comment analysis also elicited “goal setting” as a key theme. Several candidates identified future opportunities for further professional development across a broad range of areas, including policy development, finance, teaching, and strategy.

Mentoring

The influence of the NHSDA on the provision and reception of mentorship was variable. Only 27% (11/41) of the survey respondents felt they acquired mentoring skills; however, 59% (24/41) reported that they were more able to develop capabilities within their teams. Moreover, 50% (15/30) of dissertations reflected the provision of mentoring within participants’ local organizations. Following attendance at the NHSDA, some candidates were encouraged to “develop the professional training development with my [their] own teams,” with a common theme being “paying forward” the knowledge they had acquired. Furthermore, the program provided opportunities for candidates to receive mentorship or “shadowing opportunities.” One candidate who developed an interest in user-centered design “spent a really impactful afternoon with a module lead” and, subsequently, connected with designers from NHS England. This culminated in a career change to a health care–based user experience department. These experiences were significantly influenced by candidates’ work environments, with others noting they were “still alone in the organization...with little guidance from management.”

Perception of the Profession

An informatician reported, “since the course...I see myself [themselves] as a facilitator of digital transformation,” as opposed to their previous notions regarding a more technical role. This was echoed by others who perceived digital leaders as change agents: “not just somebody with the skills...but the ability to make connections to bring about transformation.” For more junior candidates, the course also helped level the hierarchy within the digital ecosystem. Having previously put “digital leaders on a pedestal” and believing that becoming one “was an unachievable target,” following the course, they believed that “it [being a digital leader] is not about having all the technical knowledge but being able to pull everyone together toward a solution.” Conversely, more established digital leaders had constructed their perceptions of digital leadership before the NHSDA, with 1 CCIO explaining, “it [the course] hasn’t changed the way I perceive what I do, it has made me more effective.”

Professional Socialization

Socialization was a key untaught component of the course. In the questionnaire, 46% (19/41) reported that the MSc program influenced their feelings of socialization within digital leadership. Inductive thematic analysis of the survey elicited “valuing collaboration” as a common theme among respondents. Many were reporting that they now found value in a “network of like-minded professionals” and wanted to “understand [their] colleagues better.” Moreover, most respondents highlighted that the program taught them how to maintain effective relationships (29/41, 71%) and inspired a shared purpose among colleagues (30/41, 73%), both facilitating a common sense of belonging. A participant suggests the “main impact of the MSc was this community of leaders who understand transformation...and share knowledge with each other.” This “collaboration is helping me [them] realise they were no different.” This network facilitated wider professional socialization by providing participants with “incredible peer support” as well as “recognition within the wider community” of the NHSDA.

Reflection

Reflection upon practice was a core facet of the dissertation, with candidates actively encouraged to examine their own practice and how it correlates with their perceptions of digital health leadership. Therefore, 80% (24/30) of the dissertations demonstrated evidence of reflective practice. This encouragement to reflect upon practice has led to several candidates reporting the academy “highlighted the positives and the negatives of my [their] leadership style,” fostering a “sort of self-awareness.” Reflection was also associated with candidates refining their perceptions of the nature of digital transformation. One CCIO from cohort 1 notes the NHSDA “makes you reflect on how we embark on this challenge of having to scale digital at pace in the context of the pandemic.”

Self-Efficacy

The development of self-efficacy was found to be a key tenet of the program, with 61% (25/41) of the respondents reporting that the course had positively increased their confidence in their role. One candidate noted that “When I [they] first went on the digital academy there was an element of impostor syndrome,”
being told, “not to think about imposters, you need to think as pioneers.” Others had reflected that the digital academy had given them “the confidence to lead in digital” and “empowerment...to recognize that I [they] have the ability to do anything I [they] put my [their] mind to.” This has led to several candidates being recognized as leaders within the wider digital health ecosystem, but not necessarily in their own organizations. One clinician noted that they “had taken up a few national unpaid roles”; however, another noted that “they [the director] was not interested...did not recognize the training we had.”

Discussion

Principal Findings

This study is one of the first to demonstrate the impact of a focused program on digital health leadership on attendees’ professional identity. The findings demonstrate that the course has a diverse range of impacts including commitment to the profession, critical thinking, goal orientation, mentoring, perception of the profession, socialization, reflection, and self-efficacy. By using a dissertation of the practice model, in which students undertake a supported digital transformation project, participants are provided with an opportunity to develop metacognitive and reflective skills. The effect of this skill development lasts beyond the course, with several participants altering their leadership style and developing more agile and collaborative approaches. Furthermore, the projects enable participants to define and attain digital goals, which may have been more difficult to define, benchmark, and achieve previously. The program reaffirmed attendees’ commitment to being or becoming a digital health leader, leading to more than half of the participants changing their roles after graduation. Among the group of individuals that changed roles, almost half noted that their experience within the NHSDA had a significant impact on this decision. In addition, the program dispelled the imposter syndrome felt by emerging leaders by increasing their confidence and a sense of professional belonging. This was facilitated by the network of alumni, which may help mitigate the organizational isolation felt by some participants.

Professional identity formation has become the focus of a diverse range of fields, including medical education [26]. Among health care professionals, studies have shown that the development of a shared core value set can have substantial benefits, including improving the well-being and resilience of physicians [27]. Professional identity’s influence in other areas has a diverse range of impacts including commitment to the profession, critical thinking, goal orientation, mentoring, perception of the profession, socialization, reflection, and self-efficacy. By using a dissertation of the practice model, in which students undertake a supported digital transformation project, participants are provided with an opportunity to develop metacognitive and reflective skills. The effect of this skill development lasts beyond the course, with several participants altering their leadership style and developing more agile and collaborative approaches. Furthermore, the projects enable participants to define and attain digital goals, which may have been more difficult to define, benchmark, and achieve previously. The program reaffirmed attendees’ commitment to being or becoming a digital health leader, leading to more than half of the participants changing their roles after graduation. Among the group of individuals that changed roles, almost half noted that their experience within the NHSDA had a significant impact on this decision. In addition, the program dispelled the imposter syndrome felt by emerging leaders by increasing their confidence and a sense of professional belonging. This was facilitated by the network of alumni, which may help mitigate the organizational isolation felt by some participants.

One of the key unintended consequences of this course has been its impact on professional socialization. Socialization is crucial for emerging learners to learn the values and beliefs necessary to succeed within their roles as well as to form a robust idea of what constitutes a digital health leader. The peer support, or “sphere of networking,” that has developed among participants has facilitated not only knowledge sharing but also a sense of community of digital health leaders. Several participants refer to a sense of confidence and validation of their identity as they were able to collaborate with recognized digital leaders. This socialization is seen in other areas of health professional development and provides a sense of “belonging,” as well as facilitating transition across clinical roles (eg, clinician to leader) [14]. This may mitigate the varying support that participants receive within their organizations.

On the other hand, few participants reported being mentored, and many participants felt unrecognized within their local institutions. Mentors are a crucial facet of identity, as it enables the observation, modeling, and imitation of leadership behaviors, as described by the social learning theory [32,33]. Consistent with previous studies, time pressures and competing demands are often barriers to mentoring in health care environments. Moreover, several participants reported a lack of recognition by their local management teams following the course. This lack of external validation as an emerging leader in digital health may have thus contributed to this shortage of mentorship opportunities. However, having engaged in the collaborative environment of the NHSDA, participants were more open to facilitating the future training of more junior members of their own teams. In addition, these local barriers may underpin the drive to find different opportunities and explain the high rates of public funding.
of role switching after graduation. Future work should look to examine these findings as well as how accreditation from courses such as the NHSDA can impact organizational buy-in.

Limitations
However, these findings must be considered within the limitations of the study. Despite using a robust approach, the respondents represent a subsection of the eligible cohorts involved. Moreover, only high-scoring dissertations were evaluated, which may have skewed our findings. However, this decision was made because scores were given based on the comprehensiveness of the write-up not the quality or results of the project. Therefore, they provided a more detailed impression of the elements of professional identity included. These selection effects were mitigated by delivering the questionnaire widely, and not all perspectives could be explored. This may affect the generalizability of the results, but it does provide a strong indication of the breadth of influences of the course. Furthermore, as previously mentioned, there is no set definition of what it is to be a digital health leader. As such, components from other contexts have been used to frame this study, which may mean that certain nuances have been omitted. Although the use of a previous extensive systematic review reduced the likelihood of this, it cannot be considered comprehensive. Furthermore, both cohorts enrolled undertook at least part of their study during the COVID-19 pandemic, in which there was a significant change in the delivery of health care and the need for digital solutions [34]. The influence of these changes on participants’ experience of the course or its impact on their professional identity cannot be ascertained. Future work should examine what constitutes a digital health leader and how this differs from health leadership more generally. This could potentially result in defining a core value set to facilitate the evaluation of digital and clinical leadership courses. This examination would need to consider the technical and nontechnical aspects of digital health leadership, as understanding both facets is essential as digital transformation continues to accelerate.

Conclusions
The increasing demand for clinical management to guide the next stages of transformation efforts requires a digitally adept corps of health leadership professionals. These digital leadership proficiencies must not only encompass technical skillsets but also include the values, judgments, and cultural beliefs about what it is to be a digital leader. The NHSDA and similar courses are likely to impact this identity formation through a broad range of effects, including socialization and professional commitment. However, further work is needed to understand what attributes are needed by a digital health leader so that training courses can be iterated and adapted. Moreover, this categorization will support the recognition of potential digital leaders who can be mentored within their local organizations, and key barriers to this progression can be overcome.

Acknowledgments
Funding for this project was provided by the Imperial Biomedical Research Centre. The Imperial Biomedical Research Centre provided the infrastructure to design and conduct the study. The funder had no role in the conduct, analysis, or dissemination of this study. Researchers were independent of funders, and all authors had full access to all the data in the study and can take responsibility for the integrity of the data.

Data Availability
The data that support the findings of this study are available from the authors, but restrictions apply to the availability of these data, which were used under license for this study and thus are not publicly available.

Authors' Contributions
AA, AS, and RCB were all involved in the study design, conduct, and data analysis. AA and RCB drafted the manuscript, with AS involved in editing and reviewing the manuscript submission. AD provided infrastructural support that enabled the study to occur and oversaw study conduct.

Conflicts of Interest
AD is the codirector of the National Health Service Digital Academy and Chair of the Health Security initiative at Flagship Pioneering UK Ltd. RCB is the Principal Teaching Fellow for the Master of Science in Digital Health Leadership and Chair of Master of Science Dissertations across the Institute of Global Health Innovation. Both authors acted independently during the conduct of this study.

Multimedia Appendix 1
Survey results and interview topic guide.

References


Abbreviations

CCIO: chief clinical informatics officer
PGDip: Postgraduate Diploma
MSc: Master of Science
NHS: National Health Service
NHSDA: National Health Service Digital Academy

©Amish Acharya, Ruth Claire Black, Alisdair Smithies, Ara Darzi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Development of a Clinical Simulation Video to Evaluate Multiple Domains of Clinical Competence: Cross-Sectional Study

Kiyoshi Shikino1,2*, MD, MHPE, PhD; Yuji Nishizaki3*, MD, MPH, PhD; Sho Fukui4, MD, MPH; Daiki Yokokawa2, MD, PhD; Yu Yamamoto5, MD; Hiroyuki Kobayashi6, MD, PhD; Taro Shimizu7, MD, MPH, PhD; Yasuhiro Tokuda8,9, MD, MPH, PhD

1Department of Community-Oriented Medical Education, Chiba University Graduate School of Medicine, Chiba, Japan
2Department of General Medicine, Chiba University Hospital, Chiba, Japan
3Division of Medical Education, Juntendo University School of Medicine, Tokyo, Japan
4Department of Emergency and General Medicine, Kyorin University, Tokyo, Japan
5Division of General Medicine, Center for Community Medicine, Jichi Medical University, Tochigi, Japan
6Department of Internal Medicine, Mito Kyodo General Hospital, Tsukuba, Japan
7Department of Diagnostic and Generalist Medicine, Dokkyo Medical University Hospital, Tochigi, Japan
8Muribushi Okinawa Center for Teaching Hospitals, Okinawa, Japan
9Tokyo Foundation for Policy Research, Tokyo, Japan
*these authors contributed equally

Corresponding Author:
Kiyoshi Shikino, MD, MHPE, PhD
Department of Community-Oriented Medical Education
Chiba University Graduate School of Medicine
1-8-1, Inohana
Chiba, 2608677
Japan
Phone: 81 43 222 7171
Email: kshikino@gmail.com

Abstract

Background: Medical students in Japan undergo a 2-year postgraduate residency program to acquire clinical knowledge and general medical skills. The General Medicine In-Training Examination (GM-ITE) assesses postgraduate residents’ clinical knowledge. A clinical simulation video (CSV) may assess learners’ interpersonal abilities.

Objective: This study aimed to evaluate the relationship between GM-ITE scores and resident physicians’ diagnostic skills by having them watch a CSV and to explore resident physicians’ perceptions of the CSV’s realism, educational value, and impact on their motivation to learn.

Methods: The participants included 56 postgraduate medical residents who took the GM-ITE between January 21 and January 28, 2021; watched the CSV; and then provided a diagnosis. The CSV and GM-ITE scores were compared, and the validity of the simulations was examined using discrimination indices, wherein ≥0.20 indicated high discriminatory power and >0.40 indicated a very good measure of the subject’s qualifications. Additionally, we administered an anonymous questionnaire to ascertain participants’ views on the realism and educational value of the CSV and its impact on their motivation to learn.

Results: Of the 56 participants, 6 (11%) provided the correct diagnosis, and all were from the second postgraduate year. All domains indicated high discriminatory power. The (anonymous) follow-up responses indicated that the CSV format was more suitable than the conventional GM-ITE for assessing clinical competence. The anonymous survey revealed that 12 (52%) participants found the CSV format more suitable than the GM-ITE for assessing clinical competence, 18 (78%) affirmed the realism of the video simulation, and 17 (74%) indicated that the experience increased their motivation to learn.

Conclusions: The findings indicated that CSV modules simulating real-world clinical examinations were successful in assessing examinees’ clinical competence across multiple domains. The study demonstrated that the CSV not only augmented the assessment of diagnostic skills but also positively impacted learners’ motivation, suggesting a multifaceted role for simulation in medical education.
Introduction

Japan’s medical schools follow a 6-year curriculum comprising 4 years of preclinical and 2 years of clinical education, after which they enter a 2-year postgraduate residency program as “postgraduate residents” or simply “residents” [1-3]. This residency enables new doctors to acquire and practice basic clinical knowledge, problem-solving, general medical and communication skills, and a professional attitude. All residents receive supervised training as they rotate through 7 specialties over the 2 years, including internal medicine, surgery, pediatrics, obstetrics and gynecology, psychiatry, emergency medicine, and community medicine. Most residents then enter specialty-based residency training.

In 2011, the nonprofit Japan Institute for Advancement of Medical Education Program (JAMEP) developed the General Medicine In-Training Examination (GM-ITE), an in-training examination for assessing the clinical knowledge of residents, similar to the US Internal Medicine Residency Examination [4]. The purpose of the GM-ITE is to elicit practical feedback on the training programs aimed at identifying improvement areas using an objective and reliable assessment of residents’ clinical knowledge [5].

The traditional assessment of clinical competencies through multiple-choice questions (MCQs), while valuable, may not encompass the full scope of a clinician’s diagnostic process in real-world practice [6]. In clinical settings, physicians must navigate through complex problem-solving and decision-making processes, often divided into domains such as leading or working diagnosis, management and treatment, hypothesis generation, problem representation, diagnostic justification, and information gathering [7]. Video simulation, as an assessment tool, can capture these nuances by providing contextualized real-world scenarios where residents must apply their knowledge dynamically, as they would in actual patient interactions [8].

Designed by a committee of experienced attending physicians organized by the JAMEP, the 2-hour GM-ITE comprises 80 MCQs covering multiple domains [9]. The scores range from 0 to 80, with higher scores indicating better performance and knowledge of internal medicine. The content and validity of each question undergo review by JAMEP’s question-development committee comprising experienced physicians from various fields, an independent peer-review committee, and examination-analysis experts [10]. The GM-ITE is not used as a pass or fail test for training advancement but only as a source of education feedback. The test is strictly voluntary, and approximately one-third of residents take the examination each year (7669 in the 2020 academic year, 6869 in the 2019 academic year, 5593 in the 2017 academic year, and 4568 in the 2016 academic year) [11,12].

An assessment of the validity of the GM-ITE [10] revealed a strong positive correlation between GM-ITE scores and scores on the Professional and Linguistic Assessments Board test, Part 1, designed to assess the depth of medical knowledge and levels of medical and communication skills [13]. In validity testing, the discrimination index (DI) indicates how well the item differentiates between students of high and low aptitude, that is, whether high-aptitude students performed better, worse, or the same as low-aptitude students [14]. Therefore, an item with a high DI is more effective in identifying respondents with adequate knowledge than an item with a low DI. The GM-ITE has indicated better discriminative power than the Professional and Linguistic Assessments Board test, Part 1 examination [10].

The JAMEP based the content of the GM-ITE on the clinical training objectives presented by Japan’s Ministry of Health, Labour and Welfare [13], which requires residents to master skills related to professionalism, physical examination and clinical procedures, and the diagnosis and treatment of common diseases. The GM-ITE shows evidence of generalization by covering 4 categories, including medical interview or professionalism (MP), clinical diagnosis (CD) consisting of symptomatology and clinical reasoning, physical examination or procedure (PP), and disease knowledge (DK). However, the relatively small number of questions in the GM-ITE provides evidence of low generalization.

Given the large number of residents taking the GM-ITE each year, using MCQs seems both expedient and appropriate when considering the viability and sustainability of the GM-ITE. However, a 2-hour test comprising only MCQs may not adequately assess the situational variations affecting clinical performance or competence in multiple domains. Therefore, this study developed a clinical simulation video (CSV) named “innovative examination” for the GM-ITE to assess residents’ clinical competency in a real-world setting using two components: (1) a high-quality CSV showing a medical interview and physical examinations with a patient and family in an emergency room and (2) follow-up questions for the residents to provide their diagnosis and recommendations. The study then evaluated the relationship between the participants’ GM-ITE and CSV innovative examination test scores by comparing their discriminative ability in each assessment domain. Therefore, this study aimed to evaluate the relationship between GM-ITE scores and resident physicians’ diagnostic skills by having them watch a CSV and to explore resident physicians’ perceptions of the CSV’s realism, educational value, and impact on their motivation to learn.
Methods

Study Design
We conducted a multicenter cross-sectional observational study in Japan.

Study Participants
The study extended an invitation to all 8526 resident physicians who took the GM-ITE in the 2021 academic year (January 21-28, 2021) to voluntarily participate in the innovative examination, and 56 residents—23 from postgraduate year (PGY) 2 and 33 from PGY 1—agreed and participated. These individuals were selected from the entire cohort of residents who took the GM-ITE. Owing to the exploratory nature of this study and the extensive distribution of the questionnaire to all eligible resident physicians, no formal sample size calculation or power analysis was performed.

Procedures

Innovative Examination Using High-Quality Patient-Simulated Video
In this study, we wrote a script depicting a simulated clinical interaction. The approximately 5-minute video ("innovative examination"), shot from a resident’s point of view, depicts a newly arrived patient and his family at an emergency room (Multimedia Appendix 1). The resident conducts a medical interview and examination, asking and answering questions, while the camera records the patient’s and family members’ verbal and nonverbal responses. Professional actors coached by the medical supervisors played the roles effectively. A professional television production company shot the video and added effects (eg, heart sounds). In total, 3 of the authors (KS, YN, and SF) and 3 JAMEP medical supervisors oversaw the video production. The study participants watched the video immediately after completing the GM-ITE. Next, they answered the CSV innovative examination questions described below.

Extended Matching Questions
We used extended matching questions that listed the patient’s symptoms to obtain up to 3 pertinent positive findings that contributed to the diagnosis (Q1 and Q2 in Textbox 1).

• Q1. Which 3 physical findings would you expect to be positive in this patient? Please choose 3 of the following:
  • Pallor of the eyelid conjunctiva
  • Pupil irregularity
  • Angry external jugular vein
  • Cervical vascular murmur
  • Thyroid gland enlargement
  • “Fixed” splitting of the second heart tone
  • Loud P2
  • Systolic murmur
  • Diastolic murmur
  • Torsion sound at the base of the lung
  • Tender points in the abdomen
  • Fresh blood in stool on rectal examination
  • Barre sign positive
  • Muscle stiffness
  • Loss of tendon reflexes

• Q2. Please state the most likely diagnosis for this patient (free text).

• Q3. Following the SBAR (situation, background, assessment, and recommendation) format, please prepare a patient handoff record for the internal medicine physician in charge of admission.
  • Q3-1. Situation (free text, 100 words maximum)
  • Q3-2. Background (free text, 100 words maximum)
  • Q3-3. Assessment (free text, 100 words maximum)
  • Q3-4. Recommendation (free text, 100 words maximum)

• Q4-1. Do you think the simulated patient-examination video was better suited to assessing your clinical competence than the traditional all-text format?
• Q4-2. Was the video simulation realistic enough for you to assess the patient?
• Q4-3. Did this experience increase your motivation to learn?

Modified Essay Questions
The third question required brief free-form answers (Q3 in Textbox 1).

Anonymous Posttest Questionnaire
After the participants completed Q1-Q3, we asked them to answer a fourth question (anonymously) to briefly describe (in writing) their experiences with the CSV innovative examination (Q4 in Textbox 1). Only 23 (41%) of the 56 participants chose to answer Q4.

Measurements
The GM-ITE uses a methodology similar to the US Internal Medicine Residency Examination [4,15,16]. The 80 questions cover 4 main categories: MP (8 questions), CD (18 questions), PP (18 questions), and DK (36 questions). We examined the validity of the GM-ITE questions using the DI φ as defined by equation 1 [17]:

\[
\phi = \frac{a - b}{\sqrt{(a + b)(c + d)}}
\]

where \(a\) is the number of correct answers in the top 25th percentile, \(b\) is the number of incorrect answers in the top 25th percentile, \(c\) is the number of correct answers in the bottom 25th percentile, and \(d\) is the number of incorrect answers in the bottom 25th percentile. The range of \(\phi\) is \(-1 \leq \phi \leq 1\). Questions are considered unreliable if this index is below 0. A DI of \(\geq 0.20\) would indicate that the question has high discriminatory power, and a DI of \(\geq 0.40\) would indicate that the question is a very good measure of the subject’s qualifications.

Statistical Analyses
We conducted these analyses using SPSS Statistics for Windows (version 26.0; IBM Corp.), following the Strengthening the Reporting of Observational Studies in Epidemiology guidelines. Two authors (KS and SF) independently assessed the answers and then discussed, identified, and agreed on them. We measured the interrater reliability with the \(\kappa\) coefficient (0.8-1.0=almost perfect, 0.6-0.8=substantial, 0.4-0.6=moderate,
and 0.2–0.4=fair) [18]. The Angoff method was used to define the cutoff for the DI calculation [19].

**Ethical Considerations**

This research was conducted in accordance with ethical standards and the principles of the Declaration of Helsinki. The ethics review board of the JAMEP, Tokyo, Japan, approved the study protocol (21-10). All participants read and signed the informed consent document before participating in the study. To ensure confidentiality, all participant data were anonymized prior to analysis. No compensation was provided to the participants for their involvement in this study. Informed consent was obtained from all participants for publication of identifying information in an online open-access publication. In accordance with ethical standards and journal policy, we have obtained explicit informed consent from all actors appearing in the video material associated with this study. The actors have acknowledged and agreed that the video will be published as part of the study’s material.

**Results**

A total of 8526 residents from 642 teaching hospitals in Japan took the GM-ITE in the 2021 academic year. Among these, 56 (23 PGY 2 and 33 PGY 1) residents also agreed to take the CSV innovative examination. The mean GM-ITE score of all 56 participants was 47.8 (SD 8.2). A DI revealed that several items had discrimination indices exceeding 0.2 (Table 1).

A total of 6 (11%) out of 56 participants answered Q2 correctly, and all the correct answers came from PGY 2 residents. The DI for the entire CSV innovative examination portion of the GM-ITE indicated high discriminatory power in all domains. Figure 1 shows the DI for the MP (8 questions) domain, with 6 innovative questions scoring a DI of ≥0.20, indicating its robustness in differentiating examinee proficiency.

Figure 2 focuses on the CD (18 questions) domain, with 5 innovative questions achieving a DI of ≥0.20, which is indicative of its strong discriminatory capability among examinees.

In Figure 3, the PP (18 questions) domain is analyzed, with 2 innovative questions achieving a DI of ≥0.20, reflecting its potential as a moderate discriminator of examinees’ understanding.

Finally, Figure 4 presents the DI for the DK (36 questions) domain, with 2 innovative questions achieving a DI of ≥0.20, reflecting its potential as a moderate discriminator of examinees’ understanding.

These figures collectively underscore the CSV innovative examination’s capacity to gauge clinical competence effectively, with each domain’s innovative question serving as a significant indicator of the examinees’ capabilities. In particular, for the innovative question Q2, a DI of ≥0.20 was found for both the total score and all 4 domains, indicating its robustness in differentiating examinee proficiency.

A total of 23 (41%) participants answered Q4, the anonymous questionnaire to assess the participants’ views on the CSV innovative examination. Regarding whether the simulated patient examination video was better suited to assessing their clinical competence than the traditional all-text format (Q4-1), 12 (52%) participants answered positively, 4 (17%) answered negatively, and 7 (30%) provided a neutral response. Regarding whether the video simulation was realistic enough for them to assess the patient (Q4-2), 18 (78%) responded affirmatively. Regarding whether the experience increased their motivation to learn, 17 (74%) responded positively.

<table>
<thead>
<tr>
<th>Domain (questions, n)</th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3-1</th>
<th>Question 3-2</th>
<th>Question 3-3</th>
<th>Question 3-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical interview or professionalism (8)</td>
<td>0.48</td>
<td>0.38</td>
<td>0.94</td>
<td>0.74</td>
<td>0.30</td>
<td>0.61</td>
</tr>
<tr>
<td>Clinical diagnosis (18)</td>
<td>0.50</td>
<td>0.40</td>
<td>0.77</td>
<td>0.56</td>
<td>0.27</td>
<td>0.18</td>
</tr>
<tr>
<td>Physical examination or procedure (18)</td>
<td>0.52</td>
<td>0.35</td>
<td>0.39</td>
<td>0.19</td>
<td>0.22</td>
<td>0.39</td>
</tr>
<tr>
<td>Disease knowledge (36)</td>
<td>–0.09</td>
<td>0.58</td>
<td>0.13</td>
<td>0.04</td>
<td>0.27</td>
<td>–0.10</td>
</tr>
<tr>
<td>Total (80)</td>
<td>0.06</td>
<td>0.47</td>
<td>0.10</td>
<td>–0.06</td>
<td>0.01</td>
<td>–0.12</td>
</tr>
<tr>
<td>Question type</td>
<td>MC&lt;sup&gt;b&lt;/sup&gt;</td>
<td>FD&lt;sup&gt;c&lt;/sup&gt;</td>
<td>FD</td>
<td>FD</td>
<td>FD</td>
<td>FD</td>
</tr>
</tbody>
</table>

<sup>a</sup>A discrimination index of ≥0.20 indicates that the question had high discriminatory power; a discrimination index of >0.40 indicates that the question was a very good measure of the participant’s qualifications.

<sup>b</sup>MC: multiple choice.

<sup>c</sup>FD: free description (<100 words).
Figure 1. DIs of the examination scores of the General Medicine In-Training Examination: medical interview or professionalism (8 questions). DI: discrimination index; Q: question.

Figure 2. DIs of the examination scores of the General Medicine In-Training Examination: clinical diagnosis (18 questions). DI: discrimination index; Q: question.
Discussion

Principal Findings

Residency is the final stage of medical education and supervised clinical practice. The traditional all-text GM-ITE was designed to elicit practical feedback on the preresidency training to identify areas of improvement by objectively assessing residents’ clinical knowledge in 4 areas: MP, CD, PP, and DK. Medical education has historically relied on MCQs to assess learning [20,21]. However, some studies have explored “context-rich” MCQs that embed test items in a clinical vignette [22,23]. This study delved beyond a written clinical vignette by creating a video simulation of a patient examination in an emergency room. The strength of ratings regarding the measures of different components of clinical reasoning indicates that although MCQs are effective in leading or working diagnosis and management and treatment, they are weak in hypothesis generation, problem representation, and diagnostic justification [7]. Conversely, it has been found that while differential diagnosis, leading or working diagnosis, diagnostic justification, and management and treatment are effective in essay style (free text), they are relatively weak in information gathering [24]. This finding suggests that CSV-based test modules could provide a more accurate measure of participants’ clinical knowledge and abilities than the GM-ITE.

Education, including medical education, has increasingly embraced computer-based testing. Today, students are accustomed to answering questions and writing essays via
In addition, medical interviews with patients and their families, examinations, the participants had the advantage of seeing and normally provided in the conventional paper-based assessments. This study is significant in that it provided "content-rich" clinical assessment and recommendations. Therefore, most participants were better able to describe the patient's situation and background than provide an descriptive response; specifically, the participants needed to name the most likely diagnosis. Two physicians (KS and SF) independently assessed the diagnoses and achieved an agreement rate of 1.00. The DI of Q2 was 0.4 or higher for symptomatology or clinical reasoning and diseases and 0.3 or higher for general theory, physical examination, and clinical techniques. The overall GM-ITE scores had a high identification index of 0.47. Specifically, the CSV innovative examination Q2’s requirement for participants to provide a definitive diagnosis allowed for a comprehensive assessment across all domains included in the GM-ITE. Furthermore, Q2 was distinguished as the sole question that demonstrated high DIs across individual disease categories. In addition, Q2 was the only question that also presented a high DI in each disease category.

CSV innovative examination Q3 required participants to provide an SBAR (situation, background, assessment, and recommendation) report using a total of 400 words or fewer. Two physicians (KS and SF) scored the responses independently and then rated each response against the scoring criteria and added them together. The agreement rate was as high as 0.92. It was observed that Q3 lowered the overall DI score to a high level in the general discussion. In other words, Q3 was easier for all the participants to answer than the other questions. For Q3-1 and Q3-2, the high discriminative ability was lowered for symptomatology and clinical reasoning. However, for each theory of disease, all the DIs were low, with some negative results. Therefore, most participants were better able to describe the patient’s situation and background than provide an assessment and recommendations.

This study is significant in that it provided “content-rich” clinical information. In addition to obtaining all the information normally provided in the conventional paper–based examinations, the participants had the advantage of seeing and hearing the various symptoms portrayed by a professional actor. In addition, medical interviews with patients and their families can reveal useful nonverbal information such as tachypnea and expressions indicating anxiety and pain levels. Gathering clinical information through diagnostic inference is critical in real-life scenarios. Participants may have performed better in certain domains covered in Q1–Q3 compared to their GM-ITE scores for the same domains owing to the CSV’s heightened sense of immediacy (seeing “real” people rather than reading about them) and the opportunity for diagnostic inferences in workplace-based assessments. This finding may indicate a development of clinical competence from the level of “knows how” to “shows” in Miller’s pyramid, which could lead to an advanced assessment in the cognitive domain.

Comparison to Prior Work

The discriminative efficacy of the CSV’s innovative examination in this study aligns with similar interventions. A study comparing simulation and video-based training for acute asthma management found that both methods significantly improved MCQ posttest scores, indicating an enhanced understanding of clinical methods [25]. Additionally, a study conducted at a university hospital in Pakistan revealed that a hybrid model combining video-based learning with simulation increased students’ confidence and performance in clinical skills. This suggests that digital and multimedia-enhanced methods may surpass traditional teaching modalities in certain aspects of medical education [26]. These comparisons underscore the potential of CSV-based assessments to provide a more nuanced and comprehensive measure of clinical competencies, potentially bridging theoretical knowledge and practical application more effectively in medical training.

Limitations

Although this study reveals important findings, it has several limitations. First, the number of participants included in the study was low. For the data to be more valid, the number of examinees needs to be increased. However, adding more participants would also increase the time and effort required to score the results. If all of the approximately 8000 examinees who took the GM-ITE completed the CSV innovative examination module, the scoring time required would be untenable, and adding more CSV-based modules would compound the problem. One way to overcome this limitation could be the use of a morphological analysis or to only score a statistically significant sampling.

Another limitation is related to the authenticity of the CSV. We created the abnormalities in the “patient,” such as the heart murmur and loud P2, by synthesizing sounds. We could not represent some aspects, such as the enhancement of systolic murmur on inspiration, and the apex beat was not clear, which might have confused the examinees. Furthermore, the time and expense involved in creating high-quality, realistic clinical cases would likely reduce the number of modules that could be used, which might enable the test takers to gain prior knowledge of the “correct” answers, therefore defeating the purpose of the test. Future research should determine the feasibility of including real cases and patients to maximize verisimilitude and reduce personnel and production expenses.
Conclusions

The findings of this study suggest that the CSV showed a high identification index for overall and multiple domains of competence in the conventional GM-ITE. The participants liked being able to “examine” the patient and receive visual and auditory clinical information, which improved their test scores. Overall, the findings showed that CSV modules simulating real-world clinical examinations assessed residents’ clinical competence successfully in multiple domains.

Acknowledgments

The authors thank the members of the Japan Institute for Advancement of Medical Education Program (JAMEP) for their valuable assistance. The JAMEP was involved in collecting and managing data as the General Medicine In-Training Examination (GM-ITE) administrative organization. It did not participate in designing and conducting the study; data analysis and interpretation; preparation, review, or approval of the paper; and the decision to submit the paper for publication. The authors also like to thank Editage for the English language review. This work was supported by the Health, Labour, and Welfare Policy Grants of Research on Region Medical (21IA2004) from the Ministry of Health, Labour and Welfare.

Data Availability

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

Authors’ Contributions

KS had full access to all the study data and took responsibility for the integrity and accuracy of the data analysis. KS, YN, SF, DY, and YT contributed to the study concept and design. HK, TS, and YY were involved in data acquisition, analysis, and interpretation. KS performed statistical analysis and wrote the paper. YN revised the content. YN and YT were involved in administrative, technical, and material support. All authors reviewed the final paper.

Authors KS (kshikino@gmail.com) and YN (ynishiza@juntendo.ac.jp) are co-corresponding authors for this article.

Conflicts of Interest

YN received an honorarium from the Japan Institute for Advancement of Medical Education Program (JAMEP) as the General Medicine In-Training Examination (GM-ITE) project manager. YT is the director of the JAMEP. HK received an honorarium from the JAMEP as a speaker for the JAMEP lecture. KS received an honorarium from the JAMEP as a reviewer of GM-ITE. KS, TS, and YY received honoraria from the JAMEP as examination preparers of GM-ITE. No other authors possess any competing interests.

Multimedia Appendix 1

Innovative examination.

References


**Abbreviations**

- **CD**: clinical diagnosis
- **CSV**: clinical simulation video
- **DI**: discrimination index
- **DK**: disease knowledge
- **GM-ITE**: General Medicine In-Training Examination
- **JAMEP**: Japan Institute for Advancement of Medical Education Program
- **MCQ**: multiple-choice question
- **MP**: medical interview or professionalism
©Kiyoshi Shikino, Yuji Nishizaki, Sho Fukui, Daiki Yokokawa, Yu Yamamoto, Hiroyuki Kobayashi, Taro Shimizu, Yasuharu Tokuda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Design and Development of Learning Management System Huemul for Teaching Fast Healthcare Interoperability Resource: Algorithm Development and Validation Study

Sergio Guinez-Molinos¹*, MSc, PhD; Sonia Espinoza²*, BEng; Jose Andrade²*, BEng; Alejandro Medina²*, BEng

¹School of Medicine, Universidad de Talca, Talca, Chile
²Interoperability Area, National Center for Health Information System, Santiago, Chile
*all authors contributed equally

Corresponding Author:
Sergio Guinez-Molinos, MSc, PhD
School of Medicine
Universidad de Talca
Avenida San Miguel #3748
Talca, 3460000
Chile
Phone: 56 996195268
Email: sguinez@utalca.cl

Abstract

Background: Interoperability between health information systems is a fundamental requirement to guarantee the continuity of health care for the population. The Fast Healthcare Interoperability Resource (FHIR) is the standard that enables the design and development of interoperable systems with broad adoption worldwide. However, FHIR training curriculums need an easily administered web-based self-learning platform with modules to create scenarios and questions that the learner answers. This paper proposes a system for teaching FHIR that automatically evaluates the answers, providing the learner with continuous feedback and progress.

Objective: We are designing and developing a learning management system for creating, applying, deploying, and automatically assessing FHIR web-based courses.

Methods: The system requirements for teaching FHIR were collected through interviews with experts involved in academic and professional FHIR activities (universities and health institutions). The interviews were semistructured, recording and documenting each meeting. In addition, we used an ad hoc instrument to register and analyze all the needs to elicit the requirements. Finally, the information obtained was triangulated with the available evidence. This analysis was carried out with Atlas-ti software. For design purposes, the requirements were divided into functional and nonfunctional. The functional requirements were (1) a test and question manager, (2) an application programming interface (API) to orchestrate components, (3) a test evaluator that automatically evaluates the responses, and (4) a client application for students. Security and usability are essential nonfunctional requirements to design functional and secure interfaces. The software development methodology was based on the traditional spiral model. The end users of the proposed system are (1) the system administrator for all technical aspects of the server, (2) the teacher designing the courses, and (3) the students interested in learning FHIR.

Results: The main result described in this work is Huemul, a learning management system for training on FHIR, which includes the following components: (1) Huemul Admin: a web application to create users, tests, and questions and define scores; (2) Huemul API: module for communication between different software components (FHIR server, client, and engine); (3) Huemul Engine: component for answers evaluation to identify differences and validate the content; and (4) Huemul Client: the web application for users to show the test and questions. Huemul was successfully implemented with 416 students associated with the 10 active courses on the platform. In addition, the teachers have created 60 tests and 695 questions. Overall, the 416 students who completed their courses rated Huemul highly.

Conclusions: Huemul is the first platform that allows the creation of courses, tests, and questions that enable the automatic evaluation and feedback of FHIR operations. Huemul has been implemented in multiple FHIR teaching scenarios for health care professionals. Professionals trained on FHIR with Huemul are leading successful national and international initiatives.

https://mededu.jmir.org/2024/1/e45413
interoperability; health information system; Health Level Seven International; HL7; Fast Healthcare Interoperability Resource; FHIR; certification; training; interoperable; e-learning; application programming interface; API

**Introduction**

A critical requirement for universal access to health is to have interconnected and interoperable health systems that guarantee effective and efficient access to quality data, strategic information, and tools for decision-making and people’s well-being [1]. One of the most relevant areas in medical informatics is the interoperability between health information systems. The interoperability eliminates duplication and errors in health data. For this reason, health informatics professionals must be educated about the benefits of interoperable systems. Therefore, strategic education on eHealth and interoperability standards is needed to enable health care professionals to make informed decisions [2].

The Fast Healthcare Interoperability Resource (FHIR) is an interoperability standard used in health information technology, introduced in 2011 by the Standard Developing Organization Health Level Seven International (HL7) [3]. FHIR is based on previous HL7 standards (HL7 versions 2 and 3 and Clinical Document Architecture) and combines their advantages with established modern web technologies such as a Representational State Transfer (REST) architecture [4], application programming interface (API), XML, JSON formats, and authorization tools (Open Authorization). The main idea behind FHIR was to build a set of resources and develop http-based REST APIs to access and use these resources. FHIR uses components called resources to access and perform operations on patient health data at the granular level [5,6].

The adoption of FHIR in health information systems by developers and companies has grown in recent years with multiple applications in various fields [5,7-9]. Thus, FHIR is positioned as an interoperability standard that is easy to understand by nontechnology professionals, with fast learning curves that minimize the development time of applications and new tools. In addition, its technological core is aligned with the latest architectures and web standards that allow the development of open APIs, which facilitates interoperability between systems [10].

Teaching and learning interoperability standards, particularly FHIR, within digital health education programs have been oriented more toward delivering content, presentations, and audiovisual material, considering the solution of practical problems separately [2]. Continuously emerging new technologies (synchronous and asynchronous) promise new and improved experiences for individual users but often bring new challenges [11].

The existing learning management systems (LMSs) are oriented to support cross-cutting activities (forums, chat, and content uploading) with content delivery (videos, documents, and links) [12] but not to evaluate REST operations for accessing and using resources. For the use of APIs, some platforms allow interaction with FHIR servers, such as Postman (Postman, Inc) or Insomnia (Kong Inc). However, they cannot create content, manage questions, automatically evaluate the response, or provide feedback but only act as an interface between the user and the FHIR server.

The configuration currently used to teach FHIR is to publish the contents in an LMS or website and, for practice, use tools such as Postman [13,14] without the possibility of having automatic feedback and correction of the activities. The results of the practical exercises must be uploaded as a document to the LMS, with written create, read, update, and delete (CRUD) operations and server response in plain text. The teacher must review them, which makes it challenging to implement workshops with many questions for large groups of students. Other websites offer the opportunity to learn FHIR with guides and theoretical content, such as Simplifier (Firely Corporation). It should be noted that Simplifier is a platform for building FHIR implementation guides. It does not claim to be an LMS or to manage courses.

There is currently no LMS for training on FHIR that allows problem-oriented assessment and practice of web-based CRUD operations. Practice is essential to learn FHIR; therefore, a problem-oriented platform is necessary, allowing the creation and administration of practical courses (where a problem is presented) with different levels of complexity and for multiple professionals (clinicians, engineers, and technicians). In addition, each course should be associated with a set of exercises, which the students must answer with CRUD operations (eg, create a patient with the data given in the description or modify the patient information with the new phone number provided). The platform should automatically evaluate these answers, and feedback should be provided to guide each question’s achievement (or nonachievement). This would help generate an extensive repository of massive web-based training programs focused on specific problems, where students must practice as requested. The lack of such platforms has motivated the interoperability team of the National Center for Health Information System (CENS) [15] to design a tool capable of automatically teaching and evaluating FHIR.

In this sense, our goal was to develop an API that allows us to integrate and communicate a set of loosely coupled modules that enable teachers to manage FHIR training programs, designing courses, questions, and scenarios. In addition, learners can interact through a web client for self-learning sessions, where the API, in conjunction with an assessment engine, provides feedback for each attempt the learner makes. This undoubtedly streamlines the self-learning process and automates the correction of hundreds of CRUD operations and the submission of learner responses within a context that the platform delivers.
The design and development of a platform called Huemul support the creation of courses associated with multiple questions (which expect a CRUD operation as an answer), automate the evaluation of the responses, and provide automatic feedback to the students in each exercise. We have also created an administrator that allows us to create and manage courses, questions, and users.

Methods

Study Design

The e-learning system requirements for teaching FHIR were collected through interviews with experts involved in academic and professional activities (universities and health institutions). The interviews were semistructured, recording and documenting each meeting. In addition, we used an ad hoc instrument to register and analyze all the needs to elicit the requirements.

The CENS academic committee, formed by 5 senior biomedical informatics researchers (3 engineers: 2 biomedical and 1 informatics and 2 medical doctors), was the initial core of experts consulted. In another focus group, engineers from the interoperability area of CENS, experts in FHIR, were consulted. They presented their requirements and needs to automate both the deployment and evaluation of the different interoperability challenges organized by CENS, where the need to register, quantify, and evaluate the hundreds of requests sent by the participants to the server was a problem when assessing their tests. These interoperability events were part of Chile’s CENS human capital training program.

Both academics and CENS engineers were interviewed with the following questions: Do you think a platform for teaching HL7 FHIR is necessary? What functions should it have? What non-functional requirements do you think are essential for the platform? For more details, see Multimedia Appendix 1.

Finally, the students (engineers from health institutions) were consulted on the platform’s functionality, modules, and usability in the first application of the pilot. A small instrument with 5 questions on a Likert scale (scale of 1-5) was applied to assess the application and the proposed modules, considering the user interface, quality of feedback, response times, quality of the content, and the response console. In addition, 2 open-ended questions were asked about the advantages and disadvantages of the platform.

The focus group sessions were transcribed, the topics of interest were categorized (user profile, usability, perceptions of use, and design), the patterns present were identified and interpreted, and the information obtained was triangulated with the available evidence. This analysis was carried out with Atlas-ti software (Scientific Software Development GmbH). With this information, the final prototype and the website for its deployment were designed.

End users are classified according to the following profiles: (1) system administrator in charge of the deployment and administration of the modules, client, and all technical aspects of the server; (2) professor who designs the course and describes the clinical context and associated questions; and (3) students in charge of accessing the client to answer questions about the course they are enrolled in.

Requirements

The system design requirements were divided into functional and nonfunctional (Textbox 1). The system development aimed to support the functional requirements to run e-learning sessions for FHIR courses. Regarding the nonfunctional requirements, security and usability are essential to design functional and secure interfaces by considering technological aspects, learner interactions, and instructional design [16,17] (Table 1). For more details, see Multimedia Appendix 1.
Textbox 1. Functional requirements to design the system for teaching FHIR (Fast Healthcare Interoperability Resource).

1. **Test and question manager:**
   - Users’ management
   - FHIR create, read, update, and delete (CRUD)-oriented test management
   - FHIR CRUD operations
   - CRUD courses
   - Create and manage a database with questions, tests, and courses

   For an FHIR test (where the context and the problem are explained), examples of questions could be:
   - Create the patient with the information given in the description
   - Create a medical encounter
   - Modify the phone number and address of the doctor
   - Delete the patient

2. **Application Programming Interface (API) for orchestrating components:**
   - Users’ authentication management
   - Call up tests and questions
   - Validate user answers
   - Save user answers
   - Execute FHIR CRUD operation on the server

3. **Test evaluator:**
   - Evaluate answers
   - Compare questions and answers
   - Build resources with the HAPI FHIR library
   - Validate resources with standard

   The expected answer should be a CRUD operation for a FHIR test (where the context and the problem are explained). For example, for the creation of a patient, the student must complete the following:
   - Method for creating a FHIR resource (post)
   - [FHIR Endpoint]/patient (URL server and resource name)
   - Patient data (JSON format; patient information)

4. **Client application:**
   - Create responsive front end
   - Communicate using the Huemul API
   - Decoupled other components
Table 1. Tools, libraries, and relation with each software component.

<table>
<thead>
<tr>
<th>Development area and tools or libraries</th>
<th>Related component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Engine</td>
</tr>
<tr>
<td>Environment</td>
<td>✓✓✓✓✓</td>
</tr>
<tr>
<td>NetBeans</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>IntelliJ CE</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Backend</td>
<td></td>
</tr>
<tr>
<td>Python</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Python 3.6</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Celery</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Django 3.1</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Django DRF 3.1</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Java</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>OpenJDK 11</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Apache Maven</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Apache Tomcat 9</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>HapiFhir 5.3</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Front end</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Bootstrap 4.3</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>jQuery 3.1.1</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Deployment</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Docker</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Docker Compose</td>
<td>✓✓✓✓</td>
</tr>
<tr>
<td>Database</td>
<td>✓✓✓✓</td>
</tr>
</tbody>
</table>

API: application programming interface.

Software Development Methodology
The development methodology was based on the traditional spiral model. The spiral development model starts with a small set of requirements and goes through each development iteration for that set of requirements. Then, the development team adds functionality for the additional requirement in ever-increasing spirals until the application is ready for the production phase [18].

Each iteration has objectives related to the evolution of the components to be developed:

1. **Modeling and management:** in the first iteration, a functional database model was generated with the objective that it can support the definition of models related to tests, users, questions, and courses and the creation of FHIR learning tests. In addition, an administration application (Huemul Admin) was created to maintain the generated models. Once the model was built, a REST API (Huemul API) was developed to consult the information.

2. **Improvements to the data model and API:** in the second iteration, improvements to the model were included with the analysis of the previous iterations, authentication and security features of the REST API, and the creation of a web client (Huemul Client) for the consumption and interaction of the REST API.

3. **Response processing and evaluation:** in the third iteration, models for response processing are included, an interface for sending responses to the web client is added, and an engine (Huemul Engine) for response evaluation is created. The administrator creates a test planning mechanism, setting start and end times.

4. **Functional improvements and feedback:** in the fourth iteration, modifications are introduced in the processing of answers, feedback in case of incorrect answers, and the enabling of a natural resource query interface.

Each developed component has a set of tools described in Table 1, the languages used are Python (Python Software Foundation) and Java (Oracle Corporation) in the backend, and all interaction between components involves using a REST API. In addition, the front end group has some traditional libraries for client development, as it uses another API to consume resources independently and does not restrict alternative clients.
Three full-time computer engineers and the leader of the CENS interoperability area worked on the platform to create the software. It took 6 months to develop the prototype and 1 month to make modifications during the pilot implementation.

**Ethical Considerations**

It should be noted that this research complied with ethical standards in accordance with the Declaration of Helsinki (updated in 2013).

**Results**

**Overview**

Huemul has 4 components that were designed and named considering the functional and nonfunctional requirements. Therefore, the following modules are necessary to develop a scalable and robust system:

1. **Huemul Admin**: web application to create users, tests, questions, and scores.
2. **Huemul API**: communication between different components of Huemul (FHIR server, client, and engine).
3. **Huemul Engine**: answers evaluation to identify differences and validate responses.
4. **Huemul Client**: web application for users to show the test and questions.

The architecture of the developed system allows for the separation into different layers. For example, the software was built under the Model-View-Controller architecture [19] to separate the views from the data model and the business logic (Figure 1). Furthermore, since usability is one of the most important nonfunctional requirements, views use web technologies, such as HTML5, JavaScript, and CSS3, to ensure access to different web browsers.

The front end can display the courses created and managed by the administration component, where the users can answer each question. In the business-oriented layer, Huemul API interconnects with the validation engine and communicates the user’s answers to this engine, which oversees validating and reviewing their structure and content. The API is Huemul’s communication core. Once a user’s response has been validated, it connects the operation with the backend (HAPI FHIR server) and communicates the result to the client.

**Figure 1.** The system architecture of Huemul with the components and their relations. API: application programming interface; FHIR: Fast Healthcare Interoperability Resource.

**Huemul Admin**

The admin component was developed as a web application to create users, tests, and questions with associated test scores. This component is decoupled from the overall system architecture, providing independence and modularity. Figure 2 shows a set of screenshots with the main functionalities of the Huemul Admin component. It shows the questions created, associated FHIR servers, tests, users, and courses. Each mentioned element can be modified and associated with generating modular courses that are easy to administer.

It is essential when creating a course to situate the clinical scenario within a context (outpatient, emergency, inpatient, and home). This will help health professionals, who are learning about interoperability, to better design the necessary resources, and CRUD operations required to solve the problems presented.
Huemul API

The core of the communication is Huemul API. This API communicates the different components of Huemul (FHIR server, client, and evaluation engine), orchestrating the whole system. An essential task of the API is communicating between the client and the evaluation engine. The test evaluation process begins when the learner sends an answer through the Huemul client application until the response is received. Specifically, the steps are as follows (Figure 3):

1. Send a request from the client: the student sends the response through the client application.
2. Internal validation: the API performs basic validations of the request sent from the client. It validates the server URL, the headers, and the body of the JSON content.
3. Engine validation: performs a full validation by comparing the answer sent by the student with the expected answer configured when creating the question.
4. Evaluation response: once all the validations have been carried out, the result is delivered, either a successful or unsuccessful comparison.
5. FHIR request: once the expected response has been validated against the one sent, if the evaluation in the engine was successful, the student’s response is sent to the corresponding FHIR server to be saved.
6. FHIR response: the FHIR server receives the request, processes it, and assigns a destination variable to the resource to identify the student who sends the response and responds to the API.
7. Build success answer: if the response from the FHIR server is successful, the API constructs the response with the summary of the validation process, evaluation, and result from the FHIR server, which will be sent to the client application.
8. Response: the API sends the answer to the client application so that the result of its submission is displayed on the screen to the learner.
Huemul Engine
This component has the function of response evaluation, for which it evaluates 2 responses, the expected response and the user’s response. The processing comprises 3 subprocesses to finally have an evaluation result that allows us to assess if the answer is correct or to assess the percentage of completeness (Figure 4).

A FHIR request, by definition, contains the following elements to be assessed:

- Base URL of the FHIR server.
- Path of the resource or query to be made to the server.
- The header of the requested content is JSON or XML.
- The body of the resource is JSON or XML format if, in case, REST methods require a body; otherwise, the body will not have information for the request.

The methods accepted to create a question are POST, PUT, GET, and DELETE.

Huemul Client
Huemul provides a web client for users, allowing them to display the test and the questions, and is the interface with the platform. For example, on the screen for sending the answer, the question statement and essential information for answering (action, precondition, expected task, etc) are presented; there is also a button to visualize the description of the scenario, and below in notifications, the platform gives feedback to the user to improve and correct the answers (Figure 5). For more details, see Multimedia Appendix 2.

When the user enters a course, the client presents the complete scenario, including information relevant to the test. Below is a
list of the exercises to be answered; each activity has an associated answer button with different colors.

- Orange button: exercise active but still needs to be answered.
- Green button: exercise with the correct answer.
- Red button: exercise with the wrong answer.

**Figure 5.** Huemul Client with a test consisting of an explanation of the scenario and associated questions. FHIR: Fast Healthcare Interoperability Resource.

**Simple Exercises on Patient Resource**

In the following exercises, you will see a compilation of basic queries, which will allow you to interact with a FHIR server in a friendly way and with the aim of creating a resource with the data provided. You will also be able to insert, modify, delete, and extract data from a FHIR server.

**Patient Data:**
- Name: Luis Gomez
- RNI: 10:010:020-3
- Gender: male
- Date of Birth: November 18, 1979
- Marriage status: Married.
- Contact number: +56923424526.

**Links of Interest:**
- [https://www.hl7.org/fhir/patient.html](https://www.hl7.org/fhir/patient.html)
- [https://www.hl7.org/fhir/summary.html](https://www.hl7.org/fhir/summary.html)
- [https://www.hl7.org/fhir.html#content]

The exercises are based on the FHIR R4 specification.

---

**Initial Evaluation of Huemul Use**

In early 2020, we conducted a pilot project in which we invited 20 health care professionals from different national institutions (10 systems development, 3 physicians, 4 computer scientists, and 3 nurses). They were students in a pilot course that presented a clinical situation and had to answer the questions through CRUD operations with HL7 FHIR. Once the course was completed, we gave them 5 questions. The questions had 5 scores according to the Likert scale for quality: 1=very poor, 2=poor, 3=fair, 4=good, and 5=excellent.

Each question focused on evaluating aspects related to the following five dimensions:

1. End-user interface: the platform is accessible and attractive for students.
2. Quality of response: feedback provided by the platform was helpful.
3. Response times: platform response times are adequate.
4. Quality of content: course description and questions are adequate.
5. Response console: response console is intuitive and easy to use.

In addition, we incorporated 2 open-ended questions that inquired about the advantages and disadvantages of the platform. The most rates of the dimensions scored on average above 4 (response times=4.9, quality of content=5, and response console=4.6). The only dimensions that did not cut above 4 on average were end-user interface and quality of feedback, with averages of 3.4 and 3.0, respectively.

This was consistent with the qualitative analysis of the open-ended questions, where students rated the content, questions, response times, and the working console positively. In general, they expressed the platform’s usefulness for self-study of FHIR. However, the usability was criticized...
Currently, Huemul has the following usage statistics:

- Users: 416 students with one or more courses in the platform.
- Courses: 10 courses.
- Tests: 60 tests.
- Questions: 695 questions (431 used and 264 unused; 572 general questions that can be used by any teacher with a Huemul account and 123 private questions).
- Response rate: 1725 (1666 completed+59 incomplete).

During the last 3 years, including the COVID-19 pandemic, 416 students have answered the same questions to evaluate the platform (with the exact 5 dimensions applied in the 2020 pilot). The evaluation has been good, with slight improvements since the pilot in dimensions 1 and 2. The same open-ended questions were applied in each course. The general comments are good or excellent, with suggestions for improvements, mainly in usability issues. The main criticisms collected in the open questions coincide with the pilot’s answers, making comments for feedback too brief and needing more helpful information to solve the exercise. Another aspect that stands out is usability, color, and font size.

Each comment has helped us to improve, incorporating a graphic designer into the team and improving the navigability of Huemul. In addition, feedback was complemented with templates of the principal associated resources that allow students to learn in a more guided way.

The preliminary impact detected is the increase in interoperability projects associated with FHIR in Chile, where the project leaders are the professionals who participated in the CENS courses with Huemul. In addition, some professionals (clinicians and engineers) were incorporated into the government to work on national strategies linked to FHIR. Other participants were recruited for medical informatics departments in hospitals (both public and private), where they led projects with FHIR.

**Discussion**

**Principal Findings**

The Huemul FHIR learning platform was designed and developed with loosely coupled components to communicate through a central API orchestrating module communication. This design was fundamental when starting to plan, considering the development of an API rather than a platform. In addition, its decoupling allows the API to interact with different technologies and with other systems and software that students can use while maintaining the independence of the components.

Integrating information dispersed in different systems is a relevant problem in health informatics. Thus, health informatics professionals must strengthen interoperability by learning standards that allow proper use. Currently, the most promising interoperability standard is FHIR. It builds on the concepts of the previous HL7 standards. The main objective of FHIR is to facilitate the implementation of solutions in various contexts: mobile apps, cloud communications, telemedicine, and medical records data sharing, among many others. Therefore, one of its main strengths is its ease of use and better learning curve compared to previous standards; this allows doctors, nurses, and engineers to work together in designing interoperable health care solutions.

To develop competencies in FHIR, Huemul has been fundamental for training professionals in Chile. The CENS [15], with its Health Information Systems (HIS) Reference Competency Model [20], has developed and used it to strengthen and generate competencies in interoperability and standards, especially with HL7 FHIR. The model proposed by CENS brings together consensual knowledge, skills, and attitudes as a reference that guides the training of excellence in biomedical informatics. Moreover, the model drives the design of undergraduate and postgraduate training curricula and establishes common training standards in the country and the region. In addition, it makes it possible to make it evident on what is expected of professionals and technicians in this sector and what is expected of them from the point of view of job opportunities or professional development.

In Chile and Latin America, there is a need for biomedical informatics professionals trained in interoperability and standards for sharing data between HIS [2]. Currently, the demand for professionals with these competencies has increased the digital gap in health and, consequently, has slowed down the changes needed to have a more connected health with robust standards, terminologies, and HIS. Huemul is available for training processes that require new ecosystems and models.

In this context, Huemul is a web application that creates users, tests, and questions to define scores and reviews them automatically in interoperability scenarios with HL7 FHIR. Huemul was the learning platform for Chile’s annual health interoperability meeting in 2020 and 2021 [21]. The interoperability meeting featured 4 sections of HL7 FHIR exercises (patient, diagnostic report, electronic medical prescription, and electronic health record), with 2 levels of complexity: introductory and intermediate. More than 100 participants each year performed hundreds of CRUD operations per exercise, which Huemul reviewed automatically. In addition, Huemul has been the official CENS platform for delivering HL7 FHIR training courses.

As a result, in the last 3 years, more than 400 technicians, engineers, and health professionals interested in learning FHIR from all over the country have been trained so far [20]. Moreover, the CENS academic team generated 10 courses with 60 associated tests. Huemul has made it possible to create a repository with more than 695 questions with different complexity levels. Each applied course has served as feedback, considering that we asked the students about the quality of our platform; considering all the dimensions exposed in the results, the users have a good evaluation of Huemul. We are still working on usability and feedback on the answers; we believe that we must improve and move forward, for example, to mobile devices and expand the content base and application areas.

Most trained professionals are leading interoperability projects with FHIR from the government, universities, and public or
private health institutions. CENS continues to support capacity building for both professionals and institutions. In this sense, Huemul is an effective tool to support practical activities, enabling the teaching of FHIR. We expect to continue advancing and complementing Huemul with new functionalities and modules in future work.

Future Work
Concerning future work, Huemul is currently in the process of redesigning for a 2.0 version that will allow us to incorporate new functionalities:

- Incorporate extensions, profiles, and extended Huemul for more search parameters. This would allow the number of questions, courses, and scenario options to be expanded as well as the complexity of the tests.
- Incorporate multiple choice and true and false questions to prepare for the HL7 FHIR certification examination. Incorporating content questions would give us a robust tool to prepare the CRUD operations in a clinical scenario and the theoretical context that will enable us to schedule examinations and certifications (eg, HL7 FHIR Proficiency examination).
- Create web-based courses with LMSs (for instance, Moodle) and Huemul. Integration with LMS platforms would extend the teaching ecosystem, incorporating content management systems, chat, forums, and all the tools with LMS.
- Incorporate other FHIR servers. Until now, Huemul has been working with HAPI FHIR, which is a complete implementation of the HL7 FHIR standard for health care interoperability in Java [22]. The advantage of having a decoupled system is the ease and modularity of its components. Huemul currently works with HAPI FHIR as a server; however, another server could be incorporated.

Another interesting aspect is evaluating and certifying interoperability levels in health information systems in a natural context [23]. Huemul could extend its applicability to other domains, for example, the assessment of HIS interoperability in hospitals, clinics, and all types of health institutions. Any modifications to its approach would be minimal, as its original 4-component structure would be maintained: Huemul Admin, Huemul API, Huemul Engine, and Huemul Client. The main changes should focus on the client-submitted request evaluation engine, broadening its focus from teaching HL7 FHIR to a more enterprise-based domain.

Considering a detailed systematic evaluation, the platform’s usability is interesting to investigate deeply. Therefore, a study design that allows the application of validated instruments and the collection of information from multiple profiles and professionals is proposed as future work.

Conclusions
Huemul is the first platform that allows the creation of courses, questions, and scenarios that enable the automatic evaluation and feedback of CRUD operations with HL7 FHIR. Huemul has been implemented and applied in multiple HL7 FHIR teaching scenarios for health care professionals. It has demonstrated its efficiency and effectiveness in courses and massive events, managing hundreds of users and evaluating thousands of answers in these 4 years of application.

Of the 416 students who were trained with Huemul, many are currently leading interoperability projects with HL7 FHIR, both in the government and the private sector, contributing to developing digital health and information systems in Chile.

Acknowledgments
This study was funded by the National Center for Health Information System (CENS) Project (CTI230006).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Huemul functional requirements.
[DOCX File, 30 KB - mededu_v10i1e45413_app1.docx ]

Multimedia Appendix 2
User manual client.
[DOCX File, 2869 KB - mededu_v10i1e45413_app2.docx ]

References


Abbreviations

API: application programming interface
CENS: National Center for Health Information System
CRUD: create, read, update, and delete
FHIR: Fast Healthcare Interoperability Resource
HIS: Health Information Systems
HL7: Health Level Seven International
LMS: learning management system
REST: Representational State Transfer
The Effects of Immersive Virtual Reality–Assisted Experiential Learning on Enhancing Empathy in Undergraduate Health Care Students Toward Older Adults With Cognitive Impairment: Multiple-Methods Study

Justina Yat Wa Liu1,2, PhD; Pui Ying Mak1, BSc; Kitty Chan1, PhD; Daphne Sze Ki Cheung1,2, PhD; Kin Cheung1, PhD; Kenneth N K Fong3, PhD; Patrick Pui Kin Kor1, PhD; Timothy Kam Hung Lai1, MSc; Tulio Maximo4, PhD

1School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
2Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
3Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
4School of Design, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

Corresponding Author:
Justina Yat Wa Liu, PhD
School of Nursing
The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom, Kowloon
Hong Kong
China (Hong Kong)
Phone: 852 27664097
Email: justina.liu@polyu.edu.hk

Abstract

Background: Immersive virtual reality (IVR)–assisted experiential learning has the potential to foster empathy among undergraduate health care students toward older adults with cognitive impairment by facilitating a sense of embodiment. However, the extent of its effectiveness, including enhancing students’ learning experiences and achieving intended learning outcomes, remains underexplored.

Objective: This study aims to evaluate the impacts of IVR-assisted experiential learning on the empathy of undergraduate health care students toward older people with cognitive impairment as the primary outcome (objective 1) and on their learning experience (objective 2) and their attainment of learning outcomes as the secondary outcomes (objective 3).

Methods: A multiple-methods design was used, which included surveys, focus groups, and a review of the students’ group assignments. Survey data were summarized using descriptive statistics, whereas paired 2-tailed t-tests were used to evaluate differences in empathy scores before and after the 2-hour IVR tutorial (objective 1). Focus groups were conducted to evaluate the impacts of IVR-assisted experiential learning on the empathy of undergraduate health care students toward older people with cognitive impairment (objective 1). Descriptive statistics obtained from surveys and thematic analyses of focus groups were used to explore the students’ learning experiences (objective 2). Thematic analysis of group assignments was conducted to identify learning outcomes (objective 3).

Results: A total of 367 undergraduate nursing and occupational therapy students were recruited via convenience sampling. There was a significant increase in the students’ empathy scores, measured using the Kiersma-Chen Empathy Scale, from 78.06 (SD 7.72) before to 81.17 (SD 8.93) after (P<.001). Students expressed high satisfaction with the IVR learning innovation, with a high satisfaction mean score of 20.68 (SD 2.55) and a high self-confidence mean score of 32.04 (SD 3.52) on the Student Satisfaction and Self-Confidence scale. Students exhibited a good sense of presence in the IVR learning environment, as reflected in the scores for adaptation (41.30, SD 6.03), interface quality (11.36, SD 3.70), involvement (62.00, SD 9.47), and sensory fidelity (31.47, SD 5.23) on the Presence Questionnaire version 2.0. In total, 3 major themes were identified from the focus groups, which involved 23 nursing students: enhanced sympathy toward older adults with cognitive impairment, improved engagement in IVR learning, and confidence in understanding the key concepts through the learning process. These themes supplement and align with the survey results. The analysis of the written assignments revealed that students attained the learning
outcomes of understanding the challenges faced by older adults with cognitive impairment, the importance of providing person-centered care, and the need for an age-friendly society.

**Conclusions:** IVR-assisted experiential learning enhances students’ knowledge and empathy in caring for older adults with cognitive impairment. These findings suggest that IVR can be a valuable tool in professional health care education.

*(JMIR Med Educ 2024;10:e48566) doi:10.2196/48566*

**KEYWORDS**

immersive virtual reality; undergraduate health care education; empathy; cognitive impairment

**Introduction**

**Background**

Empathy is a cognitive ability that involves understanding other people’s experiences, concerns, and perspectives, along with a capacity to communicate this understanding and the motivation to help others [1,2]. Showing empathy to patients, such as through active listening and self-awareness, is associated with improved patient outcomes and satisfaction [3,4]. When health care professionals understand the needs of patients, patients may feel more secure in relating their concerns to health care professionals and raising issues that worry them [5].

Although the Association of American Medical Colleges identifies empathy as an essential learning objective in health care education [6], undergraduate health care students have been found to have negative attitudes toward older people, affecting their willingness to work in this specialty [7-10]. This is especially true for older adults with cognitive impairment, about whom undergraduate health care students may hold stereotypes and whom they might socially stigmatize, leading to concerns about a possible lack of attentiveness in the provision of care to this group [11].

Empathy has been found to be positively correlated with the attitude of undergraduate health care students toward older adults and their willingness to care for them [12,13]. The most common methods for cultivating empathy in students include experiential training, didactic training, skills training, and a mixed methods approach [14]. Experiential learning is cognitively stimulating and has an impact on the entire person. It allows students to acquire knowledge, skills, and attitudes cognitively, affectively, and behaviorally [15]. Undergraduate health care students can benefit from experiential learning by considering the perspectives of the patients and experiencing them firsthand [16]. Experiential learning allows undergraduate health care students to gain more insights into how to solve the problems that older adults with cognitive impairment may encounter [17]. It is usually challenging for undergraduate health care students to understand the needs of older adults with cognitive impairment as these older adults may not be able to clearly communicate their needs [18]. However, through experiential learning, students can gain hands-on experiences that can give them a deeper knowledge and understanding of the challenges that older adults with cognitive impairment may be encountering [19].

Despite being suitable for enhancing empathy in undergraduate health care students, the various forms of conventional experiential learning, including service learning, role-play, and simulation-based workshops, have limitations in terms of replicating realistic scenarios and patients in an authentic environment. In addition, in situations in which students may become distracted, instruction from supervisors is always required [20]. For example, in role-play, not all students can immerse themselves in the role of the patient [21], affecting their learning experience. However, a new type of experiential learning delivered via immersive virtual reality (IVR) provides students with an environment that encompasses them perceptually and gives them the feeling of being within it [22]. Owing to IVR’s capacity to stimulate different senses concurrently, it is highly efficient in immersing users and generating a strong sense of presence. It is becoming more common to use IVR in health care education. However, there is a scarcity of research on such IVR experiences in an educational context [23].

IVR provides students with a realistic but safe virtual clinical environment, allowing them to gain insights into patients’ perspectives through their eyes, voices, and emotions [24]. Buchman and Henderson [25] reported that undergraduate health care students had enhanced empathy and felt a sense of realism and authenticity in the IVR experience, with empathy being the clear theme arising from the focus group analysis. Undergraduate health care students have undoubtedly also reported positive experiences with receiving different types of experiential learning other than IVR [26]. However, the sense of presence and realism generated from IVR is not possible in conventional experiential learning. IVR-assisted experiential learning is also a highly customized learning method targeted at achieving specific learning outcomes [27]. By using IVR, teachers can put undergraduate health care students in situations that are tailored to their learning needs and outcomes, whereas this level of customization may be challenging to attain in conventional experiential learning, which invariably uses a one-size-fits-all approach. Nursing students have also been found to have a higher level of engagement when taking part in IVR learning compared with their engagement with conventional learning methods, and teachers have found IVR to be helpful in compensating for the limited clinical placements available for students in hospitals [28].

Previous studies have recognized the effectiveness of IVR-assisted experiential learning in improving empathy among undergraduate health care students [29]. The Cognitive Affective Model of Immersive Learning by Makransky and Petersen [30] suggests that the mental state of perceiving a virtual self as one’s actual self with a heightened sense of embodiment refers to the sensation of possessing a virtual body. Hence, using a first-person viewpoint with a virtual environment through IVR
as a “perspective taking machine” could lead to a feeling of immersion and improve a participant’s level of embodiment, leading to an increase in empathy [31-33]. Scholars have also recommended that medical students participate in IVR experiential learning to improve their empathy before starting their clinical placement [34].

Despite previous studies, there has been little discussion on whether IVR-assisted experiential learning can enhance students’ attainment of learning outcomes such as understanding the special needs of older adults with cognitive impairment. Although there has been one study examining the improvement in the cognitive skills, such as communication competency, of multidisciplinary undergraduate and graduate health care students after an IVR simulation, its findings were based on the self-perceived evaluation of students [35]. This approach appears to lack a comparatively objective way of measuring learning outcomes, and the results of the study may be inconclusive as they may not reflect actual learning outcomes. To address this knowledge gap, it may be necessary to place more emphasis on comparatively objective assessments, such as teacher evaluations conducted according to preset assessment rubrics related to the learning outcomes.

Objectives

Therefore, this study aimed not only to evaluate the effects of IVR-assisted experiential learning on enhancing the empathy of undergraduate health care students toward older people with cognitive impairment (objective 1) but also to explore the students’ learning experiences, including “students’ satisfaction and self-confidence in learning” and “IVR fidelity” (objective 2), and their learning outcomes (objective 3) after attending the IVR-assisted experiential tutorial.

Methods

A multiple-methods design was used, which included a survey, focus groups, and student assignment reviews [36], to assess the effectiveness of the IVR-assisted experiential tutorial on students’ empathy and learning experiences and outcomes. This design produces more comprehensive findings than those obtained in single-method studies [37].

Participants

Convenience sampling was used to recruit participants for this study. Specifically, those invited to participate were undergraduate year-3 nursing students (n=267) who were taking the subject of gerontological nursing and year-3 occupational therapy (OT) students (n=100) who were taking the subject of human occupations. The nursing students were divided into 33 groups of 7 to 8 students each. They were invited to send a representative to participate in the focus groups. Ultimately, 23 group representatives participated in the focus groups. As a required learning activity, all students were obligated to attend the tutorial. However, they were given the option to join the study and complete surveys to share their learning experiences with the research team, of which 3 members (JYWL, PPKK, and KNKF) were subject lecturers. Only those who consented to join the study were included in the analysis and reporting of the results, and their anonymity was maintained in this paper.

Design of the IVR-Assisted Experiential Tutorial

Overview

To ensure that students had a solid grasp of the foundational knowledge in the subjects of gerontological nursing (for nursing students) and human occupations (for OT students), a 2-hour IVR-assisted experiential tutorial was arranged in week 7, halfway through the 13-week courses. Only the nursing students were mandated to complete and submit a group assignment within 2 weeks following the IVR tutorial. The research team developed 2 IVR games that simulated experiences commonly encountered by older adults with cognitive impairment. The first IVR game simulated a scenario in which an individual with cognitive impairment gets lost in a community setting (Figure 1). The second IVR game simulated the distorted auditory and visual perceptions commonly experienced by older adults with delirium (Figure 2). These are common challenges faced on a daily basis by older adults with cognitive impairments. These 2 IVR games were used in the 2-hour IVR-assisted experiential tutorial. Each tutorial comprised students aged between 25 and 30 years who were divided into 7 to 8 subgroups. Each subgroup underwent concurrent IVR-assisted experiential learning.

Figure 1. Scenarios simulating getting lost when looking for a supermarket as experienced by individuals with cognitive impairment.
The intended learning outcomes of the IVR-assisted experiential learning tutorial were as follows: (1) students would gain insights into the lives of older adults with cognitive impairment and their problem-solving efforts when facing daily challenges and, thus, develop empathy toward this group of older adults, (2) students would apply the skills and knowledge that they learned about common situations to propose more inclusive solutions targeted at older adults with cognitive impairment, and (3) students would be able to develop age-friendly care plans to meet the whole-person needs of older adults with cognitive impairment.

On the basis of the experiential learning model suggested by Kolb [38], 4 stages were included in the tutorial to enhance the students’ learning experiences and outcomes.

**Stage 1: Concrete Experience Through Experiential Learning**

The students’ concrete experience was obtained by exposing them to 10 to 15 minutes of IVR environments through head-mounted devices. This involved creating a realistic and immersive virtual environment that simulated a real-world experience, allowing students to engage with the internet-based environment in a meaningful way. For example, students were required to complete some daily tasks (eg, finding a supermarket) in the virtual reality (VR) environment while overwhelmed by stimuli to mimic the experiences of older people with cognitive impairment or during delirium, such as encountering confusing noises and images played through a VR head-mounted device.

**Stage 2 and 3: Reflective Observations and Abstract Conceptualizations Through Reflective and Integrative Learning During Debriefings**

Debriefing is considered an important element in experiential-based learning that reinforces and helps consolidate learning [39]. Reflective observation involves reflecting on the experience and considering what happened during the IVR simulation. The subject lecturers guided the students to reflect on and discuss the thoughts, feelings, and emotions that they experienced during the IVR-assisted experiential learning. This reflective process can help students gain insights into their own behavior and thought patterns as well as identify areas for improvement [40].

Abstract conceptualization involves interpreting and integrating the IVR experience into existing knowledge and understanding [41]. Therefore, students were motivated to reflect on and make connections between their previous experiences with older people and the insights that they gained from the IVR games. Through this process, the students showed that they were acquiring a deeper understanding of the complexities and challenges that older people with cognitive impairment face in everyday life. At the same time, students experienced the frustration and vulnerability associated with these challenges while navigating the IVR environment. The students became aware of the need for empathy, good communication, compassion, a caring and respectful attitude, and patience when working with older people with different impairments. This reflective and integrative learning approach helped cultivate empathy among the students and gave them a deeper understanding of the needs of older people.

**Stage 4: Active Experimentation by Applying the Learning in Practical Ways**

Afterward, each subtutorial group in the nursing subject was required to submit a written group report to describe the strategies (a plan) for assisting older people with cognitive impairment to remain in society. The students were expected to relate the knowledge and experiences they had gained from IVR experiential learning to the proposed strategies. They shared their strategies with their teachers and fellow students on Blackboard (a web-based education platform; Anthology Inc). The lecturers evaluated the students’ performance on this assignment based on the predeveloped rubric. This exercise in active experimentation equipped the students with the skills that they would need to work with older people and develop their advocacy roles in practice.
Outcome Measures

Empathy Toward Older Adults (Objective 1)

Students’ empathy toward older adults (objective 1) was measured using the Kiersma-Chen Empathy Scale (KCES). The 15-item KCES was developed from the theoretical perspective of empathy, which includes cognitive (ie, the ability to understand and view the world from the perspective of other people) and affective (ie, the ability to connect with the experiences or feelings of others) aspects [42]. Each item in the KCES is rated on a 7-point Likert-type scale (1=strongly disagree; 7=strongly agree). The scores on the KCES range from 15 to 105, with higher scores indicating greater empathy toward older adults. The KCES has demonstrated good test-retest reliability, with an intraclass correlation coefficient of 0.78. It correlates positively with the Jefferson Scale of Physician Empathy [43] (r=0.52) and negatively with the cynicism subscale of the Maslach Burnout Inventory (r=−0.24) [44], providing evidence of its construct validity [42]. Students were asked to complete this web-based questionnaire 1 week before the VR-assisted experiential tutorial and return the posttest questionnaire within 1 week after the tutorial.

Learning Experience (Objective 2)

The students’ experiences in learning (objective 2) with IVR-assisted experiential learning were evaluated through a posttutorial web-based survey and a focus group interview. The Student Satisfaction and Self-Confidence scale was administered after the completion of the IVR experiential tutorial. This questionnaire contains 13 items with 2 subscales (ie, satisfaction and self-confidence). Each item is rated on a 5-point Likert scale ranging from 1 (strongly disagree with the statement) to 5 (strongly agree with the statement). The scores on the satisfaction with learning scale range from 5 to 25, and the self-confidence scores range from 8 to 40, with a higher score indicating greater satisfaction and self-confidence, respectively. Both scales had high internal reliability, with a Cronbach α of .94 and .87 for the satisfaction and self-confidence scales, respectively [45].

The Presence Questionnaire version 2.0 (PQ2) was also used to evaluate the students’ sense of presence in the IVR environments (ie, IVR fidelity; objective 2) [46,47] after the IVR-assisted experiential tutorial class. The 29-item questionnaire includes 4 subscales: involvement (score range from 0 to 84), sensory fidelity (score range from 0 to 42), adaption or immersion (score range from 0 to 56), and interface quality (score range from 0 to 21), with higher scores indicating better or higher involvement, sensory fidelity, adaption or immersion, and interface quality. The students rated their experiences on a 7-point Likert scale from 1 (not at all) to 7 (completely). The PQ2 has been found to have high internal consistency, with a Cronbach α coefficient of .90, and correlates strongly with other measures of presence (r=0.78) [46].

A trained research assistant conducted 3 focus groups, with each group comprising 7 to 8 nursing students, to explore their learning experiences (objective 2) with IVR. They were asked questions such as the following: “What was your overall experience with IVR in your learning?” “How did IVR contribute to your understanding of the daily challenges of older people with cognitive impairment?” “Did you face any challenges or difficulties while using IVR for learning?” “How did IVR compare to other learning methods?” and “What suggestions do you have for improving the use of IVR in learning?” The interviews were audio recorded and then transcribed verbatim.

Learning Outcomes (Objective 3)

In this study, the impact on the students’ attainment of the learning outcomes (objective 3) referred to the students’ ability to show their understanding of the needs of older people with cognitive impairment (intended learning outcome 1) and their ability to apply this knowledge to identify inclusive strategies to help older people stay in the community (intended learning outcome 2). Only nursing students were required to complete a group assignment to describe the plan and strategies to develop age-friendly care plans to meet older adults’ needs (intended learning outcome 3). The Design of the IVR-Assisted Experiential Tutorial section provides details on the intended learning outcomes of the tutorial, and the Stage 4: Active Experimentation section provides details on the arrangement of the assignment. The group assignment was evaluated based on the assessment rubric by the lecturers of gerontological nursing (JYWL and PPKK), who were also members of the project team.

Data Analysis

The numerical data collected via the surveys were summarized as descriptive statistics using SPSS (version 27; IBM Corp) for the analysis. Simple frequencies, percentages, means, and SDs were calculated. For the pre- and posttest assessments, paired 2-tailed t tests and confidence levels were calculated to test the differences before and after the tutorial. The level of significance was set at P<.05, and all tests were 2-tailed.

The text data collected through focus groups to identify the students’ learning experiences were analyzed using descriptive thematic analysis. To identify the students’ achievement of the learning outcomes, their written assignments were also analyzed using a descriptive thematic analysis. In contrast to other similar approaches, in thematic analysis, there is no commitment to a specific theoretical framework; therefore, a thematic analysis can be used between various theoretical frameworks. Thus, it is a more accessible and flexible form of analysis. What researchers do with the themes once they are uncovered will differ based on the aim of the research and the process of analysis [48]. In total, 2 researchers (JYWL and PPKK) read the students’ written assignments and independently identified codes from them. Codes with similar content were grouped together to form subthemes. The subthemes were then categorized into themes. Another researcher (KC) reviewed the codes, subthemes, and themes, and any discrepancies were resolved through discussion to achieve a consensus.

Ethical Considerations

This study was approved by the Human Subjects Ethics Application Review System of the Hong Kong Polytechnic University (HSEARS20200423001) and conducted between June 2021 and May 2022. It was carried out in accordance with...
the Declaration of Helsinki. This included but was not limited to guaranteeing the anonymity of participants and obtaining the informed consent of the participating students. The participation of the students was voluntary, and their academic results were not affected by their decision to participate in the study.

Results

Overview

Of the 367 students who were enrolled in the 2 subjects, 93.7% (344/367) consented to join the study, of whom 75.6% (260/344) were nursing students and 24.4% (84/344) were OT students. They completed and returned the pre- and posttest surveys with an overall response rate of 93.7% (344/367). Most participating students were female (256/344, 74.4%), 23.3% (80/344) were male, and 2.3% (8/344) did not report their gender. Their ages ranged from 18 to 24 years.

We invited all 33 subgroups from the nursing subject to send 1 representative to join the focus groups. Eventually, 23 group representatives (a response rate of 23/33, 70%) participated in the focus groups, of whom 16 (70%) were female students. The participants were assigned to 1 of the 3 focus groups, with each group comprising 7 to 8 students to facilitate in-depth group discussions.

Empathy Toward Older Adults (Objective 1)

Participating students showed moderate empathy toward older people, as reflected by a KCES score of 78.06 (SD 7.72) out of 105 before the IVR-assisted experiential tutorial. After completing the tutorial, this score increased to 81.17 (SD 8.93). The results of the paired-sample 2-tailed t test showed a significant increase in the mean score from before to after the tutorial (t_{304}=3.95; P<.001; Table 1). A further subgroup analysis was conducted, and a significant difference was found in the results between the nursing and OT students in KCES scores. There was a significant improvement in KCES scores among the nursing students but a decreasing trend among the OT students (Multimedia Appendix 1).

<table>
<thead>
<tr>
<th>Question</th>
<th>Before, mean score (SD)</th>
<th>After, mean score (SD)</th>
<th>t statistic</th>
<th>Mean difference (SD)</th>
<th>t test (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. It is necessary for a health care practitioner to be able to comprehend someone else’s experiences.</td>
<td>5.80 (0.89)</td>
<td>5.83 (0.91)</td>
<td>0.03 (1.11)</td>
<td>0.46 (304)</td>
<td>.64</td>
<td></td>
</tr>
<tr>
<td>2. I am able to express my understanding of someone’s feelings.</td>
<td>5.38 (0.92)</td>
<td>5.61 (0.91)</td>
<td>0.23 (1.04)</td>
<td>3.91 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>3. I am able to comprehend someone else’s experiences.</td>
<td>5.33 (0.85)</td>
<td>5.65 (0.88)</td>
<td>0.32 (1.06)</td>
<td>5.32 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>4. I will not allow myself to be influenced by someone’s feelings when determining the best treatment(^a).</td>
<td>4.62 (1.27)</td>
<td>4.65 (1.43)</td>
<td>0.62 (2.28)</td>
<td>4.88 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>5. It is necessary for a health care practitioner to be able to express an understanding of someone’s feelings.</td>
<td>5.77 (0.85)</td>
<td>5.85 (0.74)</td>
<td>0.79 (0.93)</td>
<td>1.48 (304)</td>
<td>.14</td>
<td></td>
</tr>
<tr>
<td>6. It is necessary for a health care practitioner to be able to value someone else’s point of view.</td>
<td>5.80 (0.88)</td>
<td>5.93 (0.82)</td>
<td>0.13 (1.04)</td>
<td>2.10 (304)</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>7. I believe that caring is essential to building a strong relationship with patients.</td>
<td>6.06 (0.79)</td>
<td>6.03 (0.82)</td>
<td>0.03 (0.89)</td>
<td>0.58 (304)</td>
<td>.56</td>
<td></td>
</tr>
<tr>
<td>8. I am able to view the world from another person’s perspective.</td>
<td>5.34 (0.94)</td>
<td>5.69 (0.85)</td>
<td>0.35 (1.15)</td>
<td>5.26 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>9. Considering someone’s feelings is not necessary to provide patient-centered care(^a).</td>
<td>3.15 (1.79)</td>
<td>3.78 (2.11)</td>
<td>0.64 (2.28)</td>
<td>4.88 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>10. I am able to value someone else’s point of view.</td>
<td>5.43 (0.87)</td>
<td>5.70 (0.84)</td>
<td>0.27 (1.05)</td>
<td>4.52 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>11. I have difficulty identifying with some else’s feelings(^a).</td>
<td>3.51 (1.43)</td>
<td>3.98 (1.66)</td>
<td>0.47 (1.88)</td>
<td>4.38 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>12. To build a strong relationship with patients, it is essential for a health care practitioner to be caring.</td>
<td>5.81 (0.88)</td>
<td>5.93 (0.82)</td>
<td>0.12 (1.02)</td>
<td>2.03 (304)</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>13. It is necessary for a health care practitioner to identify with someone else’s feelings.</td>
<td>5.73 (0.87)</td>
<td>5.94 (0.79)</td>
<td>0.20 (0.94)</td>
<td>3.77 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>14. It is necessary for a health care practitioner to be able to view the world from another person’s perspective.</td>
<td>5.69 (0.87)</td>
<td>5.90 (0.84)</td>
<td>0.21 (0.92)</td>
<td>4.03 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>15. A health care practitioner should not be influenced by someone’s feelings when determining the best treatment(^a).</td>
<td>4.81 (1.36)</td>
<td>4.70 (1.60)</td>
<td>0.11 (1.59)</td>
<td>1.15 (304)</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>Total KCES</td>
<td>78.06 (7.72)</td>
<td>81.17 (8.93)</td>
<td>3.11 (0.523)</td>
<td>3.95 (304)</td>
<td>&lt;.001</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Items with negative wordings are scored in reverse.
Learning Experience (Objective 2)

Students’ Satisfaction and Self-Confidence in Learning

Students were satisfied with the current learning innovation, as reflected by a high satisfaction mean score of 20.68 (SD 2.55) out of 25. For example, 92.7% (319/344) of the students agreed or strongly agreed that IVR-assisted experiential learning was suitable for the way they learned (item 5). The same percentage of students agreed or strongly agreed that the IVR learning experience provided an alternative learning experience to promote their learning interests (item 2). A total of 91.6% (315/344) of the students agreed or strongly agreed that the IVR simulation was motivating and helped them learn better (item 4). They also showed a high level of self-confidence in their IVR experiential learning, with a mean score of 32.04 (SD 3.52) out of 40. Approximately 85.5% (294/344) of the students agreed or strongly agreed that they were confident that they would obtain the necessary skills and knowledge through learning with the IVR simulation (items 6-8). A total of 95.1% (327/344) of the participants agreed or strongly agreed that students should take responsibility for their learning (items 10-11; Table 2).

Table 2. The findings of the Student Satisfaction and Self-Confidence scale (n=344).

<table>
<thead>
<tr>
<th>Item</th>
<th>Participants, n (%)</th>
<th>Strongly disagree (1)</th>
<th>Disagree (2)</th>
<th>Undecided (3)</th>
<th>Agree (4)</th>
<th>Strongly agree (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Satisfaction with the current learning subscale</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. The teaching methods used in the IVR&lt;sup&gt;a&lt;/sup&gt; simulation were helpful and effective.</td>
<td>2 (0.6)</td>
<td>4 (1.2)</td>
<td>20 (5.8)</td>
<td>242 (70.4)</td>
<td>76 (22.1)</td>
<td></td>
</tr>
<tr>
<td>2. The IVR simulation provided me with a variety of learning materials and activities to promote my learning curriculum.</td>
<td>1 (0.3)</td>
<td>5 (1.5)</td>
<td>19 (5.5)</td>
<td>244 (70.9)</td>
<td>75 (21.8)</td>
<td></td>
</tr>
<tr>
<td>3. I enjoyed how my instructor taught the IVR simulation.</td>
<td>1 (0.3)</td>
<td>4 (1.2)</td>
<td>25 (7.3)</td>
<td>233 (67.7)</td>
<td>81 (23.5)</td>
<td></td>
</tr>
<tr>
<td>4. The teaching materials used in this IVR simulation were motivating and helped me to learn.</td>
<td>1 (0.3)</td>
<td>5 (1.5)</td>
<td>23 (6.7)</td>
<td>232 (67.4)</td>
<td>83 (24.1)</td>
<td></td>
</tr>
<tr>
<td>5. The way my instructor taught the IVR simulation was suitable to the way I learn.</td>
<td>1 (0.3)</td>
<td>3 (0.9)</td>
<td>21 (6.1)</td>
<td>246 (71.5)</td>
<td>73 (21.2)</td>
<td></td>
</tr>
<tr>
<td><strong>Self-confidence in learning subscale</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. I am confident that I am mastering the content of the IVR simulation activity that my instructor presented to me.</td>
<td>1 (0.3)</td>
<td>7 (2)</td>
<td>45 (13.1)</td>
<td>234 (68)</td>
<td>57 (16.6)</td>
<td></td>
</tr>
<tr>
<td>7. I am confident that this simulation covered critical content necessary for the mastery of the curriculum.</td>
<td>1 (0.3)</td>
<td>6 (1.7)</td>
<td>47 (13.7)</td>
<td>239 (69.5)</td>
<td>51 (14.8)</td>
<td></td>
</tr>
<tr>
<td>8. I am confident that I am developing the skills and obtaining the required knowledge from this simulation to perform necessary tasks in a clinical setting.</td>
<td>1 (0.3)</td>
<td>11 (3.2)</td>
<td>38 (11)</td>
<td>246 (71.5)</td>
<td>48 (14)</td>
<td></td>
</tr>
<tr>
<td>9. My instructors used helpful resources to teach the simulation.</td>
<td>1 (0.3)</td>
<td>5 (1.5)</td>
<td>17 (4.9)</td>
<td>244 (70.9)</td>
<td>77 (22.4)</td>
<td></td>
</tr>
<tr>
<td>10. It is my responsibility as the student to learn what I need to know from this IVR simulation activity.</td>
<td>1 (0.3)</td>
<td>1 (0.3)</td>
<td>15 (4.4)</td>
<td>260 (75.6)</td>
<td>67 (19.5)</td>
<td></td>
</tr>
<tr>
<td>11. I know how to get help when I do not understand the concepts covered in the simulation.</td>
<td>1 (0.3)</td>
<td>6 (1.7)</td>
<td>28 (8.1)</td>
<td>254 (73.8)</td>
<td>55 (16)</td>
<td></td>
</tr>
<tr>
<td>12. I know how to use simulation activities to learn critical aspects of these skills.</td>
<td>1 (0.3)</td>
<td>5 (1.5)</td>
<td>30 (8.7)</td>
<td>246 (71.5)</td>
<td>62 (18)</td>
<td></td>
</tr>
<tr>
<td>13. It is the instructor’s responsibility to tell me what I need to learn of the simulation activity content during class time.</td>
<td>1 (0.3)</td>
<td>27 (7.8)</td>
<td>89 (25.9)</td>
<td>187 (54.4)</td>
<td>40 (11.6)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>IVR: immersive virtual reality.

IVR Fidelity

IVR fidelity was measured using the PQ2. The results showed that students developed a good sense of presence in the IVR learning environment, as seen in their scores on adaptation (mean 41.30, SD 6.03 out of 56), interface quality (mean 11.36, SD 3.70 out of 21), involvement (mean 62.0, SD 9.47 out of...
84), and sensory fidelity (mean 31.47, SD 5.23 out of 42) (Table 3).

On the basis of the focus group discussions with the students about their experiences of experiential learning with IVR, 4 themes were identified: enhanced sympathetic feeling toward older adults with cognitive impairment, improved engagement in IVR learning, confidence in understanding key concepts in the IVR experiential learning process, and limitations of IVR technology.
Table 3. The findings of the Presence Questionnaire version 2.0 (n=344).

<table>
<thead>
<tr>
<th>Item</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Involvement</strong></td>
<td></td>
</tr>
<tr>
<td>1. How much were you able to control events?</td>
<td>3 (0.9) 4 (1.2) 3 (0.9) 45 (13.1) 73 (21.2) 156 (45.3) 60 (17.4)</td>
</tr>
<tr>
<td>2. How responsive was the environment to actions that you initiated (or performed)?</td>
<td>1 (0.3) 5 (1.5) 9 (2.6) 66 (19.2) 116 (33.7) 107 (31.1) 40 (11.6)</td>
</tr>
<tr>
<td>3. How natural did your interactions with the IVR environment seem?</td>
<td>8 (2.3) 6 (1.7) 23 (6.7) 64 (18.6) 121 (35.2) 92 (26.7) 30 (8.7)</td>
</tr>
<tr>
<td>4. How much did the visual aspects of the IVR environment involve you?</td>
<td>1 (0.3) 1 (0.3) 11 (3.2) 43 (12.5) 80 (23.3) 155 (45.1) 53 (15.4)</td>
</tr>
<tr>
<td>6. How natural was the mechanism that controlled movement through the environment?</td>
<td>6 (1.7) 7 (2) 26 (7.6) 48 (14) 134 (39) 99 (28.8) 24 (7)</td>
</tr>
<tr>
<td>7. How compelling was your sense of objects moving through space?</td>
<td>2 (0.6) 3 (0.9) 10 (2.9) 55 (16) 133 (38.7) 110 (32) 31 (9)</td>
</tr>
<tr>
<td>8. How much did your experiences in the virtual environment seem to be consistent with your real-world experiences?</td>
<td>14 (4.1) 13 (3.8) 24 (7) 69 (20.1) 109 (31.7) 93 (27) 22 (6.4)</td>
</tr>
<tr>
<td>10. How completely were you able to actively survey or search the IVR environment using vision?</td>
<td>1 (0.3) 3 (0.9) 6 (1.7) 41 (11.9) 132 (38.4) 121 (35.2) 40 (11.6)</td>
</tr>
<tr>
<td>14. How compelling was your sense of moving around inside the virtual environment?</td>
<td>1 (0.3) 4 (1.2) 11 (3.2) 60 (17.4) 140 (40.7) 96 (27.9) 32 (9.3)</td>
</tr>
<tr>
<td>17. How well could you move or manipulate objects in the virtual environment?</td>
<td>12 (3.5) 2 (0.6) 22 (6.4) 67 (19.5) 116 (33.7) 100 (29.1) 25 (7.3)</td>
</tr>
<tr>
<td>18. How involved were you in the virtual environment experience?</td>
<td>3 (0.9) 5 (1.5) 6 (1.7) 42 (12.2) 115 (33.4) 124 (36) 49 (14.2)</td>
</tr>
<tr>
<td>26. How easy was it to identify objects through physical interaction (eg, touching an object, walking over a surface, or bumping into a wall or object)?</td>
<td>10 (2.9) 11 (3.2) 24 (7) 91 (26.5) 126 (36.6) 58 (16.9) 24 (7)</td>
</tr>
<tr>
<td><strong>Sensory fidelity</strong></td>
<td></td>
</tr>
<tr>
<td>5. How much did the auditory aspects of the IVR environment involve you?</td>
<td>5 (1.5) 3 (0.9) 18 (5.2) 49 (14.2) 103 (29.9) 118 (34.3) 48 (14)</td>
</tr>
<tr>
<td>11. How well could you identify sounds?</td>
<td>3 (0.9) 4 (1.2) 8 (2.3) 43 (12.5) 107 (31.1) 127 (36.9) 52 (15.1)</td>
</tr>
<tr>
<td>12. How well could you localize sounds?</td>
<td>3 (0.9) 6 (1.7) 11 (3.2) 50 (14.5) 118 (34.3) 114 (33.1) 42 (12.2)</td>
</tr>
<tr>
<td>13. How well could you actively survey or search the virtual environment using touch?</td>
<td>10 (2.9) 9 (2.6) 21 (6.1) 57 (16.6) 120 (34.9) 98 (28.5) 29 (8.4)</td>
</tr>
<tr>
<td>15. How closely were you able to examine objects?</td>
<td>2 (0.6) 4 (1.2) 17 (4.9) 59 (17.2) 129 (37.5) 102 (29.7) 31 (9)</td>
</tr>
<tr>
<td>16. How well could you examine objects from multiple viewpoints?</td>
<td>1 (0.3) 2 (0.6) 12 (3.5) 64 (18.6) 118 (34.3) 116 (33.7) 31 (9)</td>
</tr>
<tr>
<td><strong>Adaption or immersion</strong></td>
<td></td>
</tr>
<tr>
<td>9. Were you able to anticipate what would happen next in response to the actions that you performed?</td>
<td>4 (1.2) 9 (2.6) 39 (11.3) 63 (18.3) 116 (33.7) 88 (25.6) 25 (7.3)</td>
</tr>
<tr>
<td>20. How quickly did you adjust to the virtual environment experience?</td>
<td>3 (0.9) 2 (0.6) 11 (3.2) 41 (11.9) 133 (38.7) 86 (25) 68 (19.8)</td>
</tr>
<tr>
<td>21. How proficient in moving and interacting with the virtual environment did you feel at the end of the experience?</td>
<td>2 (0.6) 11 (3.2) 34 (9.9) 136 (39.5) 125 (36.3) 36 (10.5) 0 (0)</td>
</tr>
<tr>
<td>24. How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities?</td>
<td>0 (0) 1 (0.3) 4 (1.2) 59 (17.2) 131 (38.1) 118 (34.3) 31 (9)</td>
</tr>
<tr>
<td>Item</td>
<td>Participants, n (%)</td>
</tr>
<tr>
<td>---------------------------------------------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>25. How completely were your senses engaged in this experience?</td>
<td></td>
</tr>
<tr>
<td>27. Were there moments during the virtual environment experience when you felt completely focused on the task or environment?</td>
<td></td>
</tr>
<tr>
<td>28. How easily did you adjust to the control devices used to interact with the virtual environment?</td>
<td></td>
</tr>
<tr>
<td>29. Was the information provided through different senses in the virtual environment (eg, vision, hearing, touch) consistent?</td>
<td></td>
</tr>
</tbody>
</table>

**Interface quality**

<table>
<thead>
<tr>
<th>Item</th>
<th>Participants, n (%)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>19. How much delay did you experience between your actions and the expected outcomes?</td>
<td></td>
<td>7 (2)</td>
<td>32 (9.3)</td>
<td>64 (18.6)</td>
<td>102 (29.7)</td>
<td>57 (16.6)</td>
<td>58 (16.9)</td>
<td>24 (7)</td>
</tr>
<tr>
<td>22. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?</td>
<td></td>
<td>17 (4.9)</td>
<td>79 (23)</td>
<td>66 (19.2)</td>
<td>96 (27.9)</td>
<td>44 (12.8)</td>
<td>24 (7)</td>
<td>18 (5.2)</td>
</tr>
<tr>
<td>23. How much did the control devices interfere with the performance of assigned tasks or with other activities?</td>
<td></td>
<td>21 (6.1)</td>
<td>74 (21.5)</td>
<td>106 (30.8)</td>
<td>73 (21.2)</td>
<td>37 (10.8)</td>
<td>15 (4.4)</td>
<td>18 (5.2)</td>
</tr>
</tbody>
</table>

*IVR: immersive virtual reality.
Reverse items.*

**Enhanced Sympathetic Feelings Toward Older Adults With Cognitive Impairment**

All participants in the focus group were impressed by the authenticity of the IVR games, which allowed them to experience the daily challenges faced by older people with cognitive impairment. One student remarked the following:

*The IVR experience allowed me to see the world from the perspective of an older person with cognitive impairment who was getting lost. This experience helped me to better understand the confusion and disorientation that older people may face, which in turn helped me to be more empathetic and compassionate toward them.*

Another student added the following:

*This VR experience was so lifelike that it helped me to empathize with their (older people with cognitive impairment) situation and understand their needs better.*

**Improved Engagement in IVR Learning**

Most participants in the focus groups said that IVR helped them stay engaged and interested in the learning process, which could sometimes be challenging in traditional classroom settings. One student said the following:

*With IVR, I was able to experience the daily challenges of older adults with cognitive impairment, which made the learning process more exciting and engaging than conventional teaching methods. With this firsthand experience, I am motivated to learn and identify strategies to help them (older adults) overcome those challenges.*

**Confidence in Understanding Key Concepts in the IVR Experiential Learning Process**

Students also showed confidence in their learning with IVR. They stated that learning with IVR improved their memory retention by providing a more realistic and memorable learning experience. One student commented the following:

*The IVR game of delirium was a great way to simulate the condition and learn how to manage it (delirium in patients). It gave me the confidence to recognize and manage delirium in a real-life situation.*

This sentiment was echoed by another student, who said the following:

*The “get lost” game made me realize the importance of taking extra precautions to ensure the safety of older people with cognitive impairment. Overall, these experiences allowed me to develop a deeper understanding of the challenges associated with caring for them, which gives me more confidence in my ability to provide effective care to them.*

**Limitations of IVR Technology**

Although IVR offered a unique and engaging learning experience for students, technical issues such as equipment malfunctions and slow processing times could limit the effectiveness of the IVR learning experience. One student stated the following:

*I encountered some technical issues during the IVR experience, which interrupted the flow of the scenario.*
and disrupted my immersion in the experience. It was frustrating, and I felt like I missed out on some important learning opportunities as a result.

Another student added the following:

The VR headset was heavy and its size needed to be adjusted continually to fit my head, making it difficult to fully immerse myself in the scenario. I found it challenging to stay focused and engaged during the entire experience.

**Learning Outcomes (Objective 3)**

**Overview**

To understand the students’ attainment of the 3 learning outcomes after completing the IVR-assisted experiential tutorial, we conducted thematic analyses of the group written assignments. The analysis was based on 33 group assignments from the nursing students. In total, 3 themes were identified.

**Understanding the Challenges Faced by Older People With Cognitive Impairment**

The analysis of the students’ written assignments indicated that they had developed a basic understanding of the challenges faced by older people with cognitive impairment. For example, one group report stated the following:

The psychological well-being of older people would be negatively influenced due to their hallucinations. It is because restlessness and agitation would be provoked by the experiences of distorted images and sounds. The situations may happen at any time, which gives the older people much mental stress.

Another statement also said the following:

Their quality of life would be seriously affected since their cognitive functions are impaired, lowering their independence in daily living. To prevent themselves from making mistakes, they (older adults) may withdraw from society or stop doing things that they used to do. Therefore, some older adults may suffer from depression and become socially isolated due to cognitive decline.

**Person-Centered Care**

This care approach was mentioned consistently in group assignments. One report stated the following:

Person-centered care is essential to ensure that older people with cognitive impairment receive care that is tailored to their unique needs and preferences. “Effective communication,” “family involvement,” and “supportive care with patience” were 3 critical aspects of person-centered care that were frequently discussed in the assignments:

Effective communication is a key component in person-centered care to ensure that this vulnerable group can express their needs and preferences so that the care can be tailored for them.

They also mentioned the need for family members to be included in the care planning and decision-making process. One group wrote the following:

Family members play a critical role in providing support and care to older people with cognitive impairment. This is particularly the case during delirium.

Their involvement can promote continuity of care and provide emotional support to their families with cognitive impairment, especially when they are in a distressing situation, such as delirium.

The need to be supportive was stated frequently in the written assignments. For example, one report stated the following:

As nurses, we need to provide support to individuals with cognitive impairment to promote their independence and autonomy. In order to empower them to be able to continue living their life with dignity, we should give them various forms of support.

**Creation of an Age-Friendly Society**

It was stated that this is an essential strategy to enable older people with cognitive impairment to stay in the community with dignity for as long as possible. In a written report, students recognized the need to reduce the stigma surrounding cognitive impairment and stated the following:

We need to raise awareness and educate people about the common daily challenges faced by older people with cognitive impairment to eliminate negative stereotypes and improve social inclusion for them.

Students also became aware of the importance of social inclusion in creating an age-friendly society, stating the following:

We need to create a supportive and inclusive environment that recognizes the unique needs of individuals with cognitive impairment.

They also suggested some concrete community-based services and support to enable this segment of the population to remain in their community for as long as possible. One group wrote the following:

Community-based services, such as transportation, social activities, and assistive technologies, can help them to stay connected and engaged in their communities.

Another group echoed this with the following suggestion:

Provide more community activities to enhance their interaction with the society, which can help the older adults expand their social circle to reduce the rate of deterioration of their cognitive function.

**Discussion**

**Principal Findings**

The results suggest that IVR-assisted experiential learning is effective in enhancing empathy toward older people among undergraduate nursing and OT students, as reflected in their higher scores on the KCES after the IVR simulation. The
students reported a high level of satisfaction with the IVR learning experience, citing its suitability, ability to motivate, and innovativeness in the self-administered survey. In addition, the findings from the survey suggest that the students experienced a strong sense of presence in the IVR learning environment, enabling them to gain a deeper understanding of the challenges involved in caring for older adults with cognitive impairment. In total, 3 major themes were identified from the focus groups with 23 nursing students: enhanced sympathetic feelings toward older adults with cognitive impairment, improved engagement in IVR learning, and confidence in understanding the key concepts through the learning process.

The thematic findings supplement and are in line with the results from the survey. The analysis of the written assignments showed that the students attained the learning outcomes of understanding the challenges faced by older people with cognitive impairment, the importance of providing person-centered care, and the need to create an age-friendly society.

These findings are consistent with those of previous studies that demonstrated the effectiveness of IVR as a mode of experiential learning to enhance the empathy of students toward older adults [49,50]. However, previous studies have mainly measured changes in students’ level of empathy using questionnaires without exploring the underlying reasons.

**Empathy Toward Older Adults and Learning Experience**

Our survey findings for objectives 1 and 2 are consistent with the insights gained from the focus groups. For example, the PQ2 scores indicated that the students felt a strong sense of presence in the IVR environment, which was also reflected in their comments during the focus groups. Participants in the focus groups mentioned that the authentic IVR games allowed them to better understand and empathize with the daily challenges faced by older people with cognitive impairment, which may have contributed to the significant increase in empathy toward older adults reflected in the KCES scores. Furthermore, both the surveys and focus groups revealed that students were satisfied with the IVR-assisted experiential learning and felt confident in their ability to understand the key concepts through this approach. These consistent findings across multiple data sources provide strong evidence to suggest the effectiveness of IVR-assisted learning in enhancing students’ empathy and understanding of key concepts as well as their satisfaction with the IVR teaching approach. Compared with conventional teaching methods, IVR creates a sense of presence and provides an excellent medium for experiencing alternative points of view, allowing undergraduate health care students to virtually “step into the shoes of older adults” [23]. The hands-on experiences provided by IVR enable students to gain a deeper understanding and knowledge of the challenges that older adults with cognitive impairment may encounter [19].

The findings of this study suggest that IVR can promote positive learning experiences, including increased satisfaction, self-confidence, self-assessed competency, self-efficacy, and enjoyment among undergraduate health care students [51]. This evidence is consistent with the positive learning experiences identified in this study based on both quantitative and qualitative data. In addition, IVR facilitates a constructivist approach to education that emphasizes active participation in the learning process rather than the passive receipt of information [52]. That was why, in the focus groups, students stated that they experienced improved engagement with this innovative learning approach. It provides active and constructivist learning and increases students’ engagement in their learning, leading to an increase in the frequency of authentic learning experiences. Being engaged encourages students to become aware of learning concepts such as empathy and other soft skills needed to care for older adults.

The subgroup analysis revealed a notable enhancement in KCES scores among nursing students in contrast to a declining trend among OT students. As the aim of this study was not to draw comparisons between these 2 student groups but rather to evaluate overall empathy levels among nursing and OT students, we are unable to explain the reasons for these differences. This discrepancy could potentially be attributed to the non–discipline-specific design of the intervention, which may have been more beneficial to nursing students than to OT students.

**Learning Outcomes**

Apart from enhancing empathetic experiences, an analysis of the students’ group assignments in this study revealed 3 major themes related to their learning outcomes [53]. These findings indicate that the students improved their understanding of the challenges faced by older people with cognitive impairment. Consequently, nursing students recognized the importance of person-centered care for this population, including effective communication, family involvement, and supportive care. Finally, the students highlighted the need to create an age-friendly society by reducing stigma, promoting social inclusion, and providing community-based services and support.

**Implications**

By improving empathy levels through IVR experiential learning, students become more capable of comprehending needs and experiences from the perspective of the patients. The empathetic response of the students can provide insights into how newly acquired knowledge of the lived experiences of older adults with cognitive impairment can be used to enhance the quality of life of these older adults [54]. In this way, students will be better equipped to develop individualized care plans tailored to the specific needs of patients [55]. IVR experiential learning also inspires students to adopt a holistic approach when providing care to older people with cognitive impairment, recognizing the significance of social and environmental factors in their care plans [56].

**Limitations and Challenges of IVR Learning**

Although IVR-assisted experiential learning has shown positive results in enhancing health care education, it is important to acknowledge the limitations and challenges associated with adopting this technology in teaching. Technical issues such as equipment malfunctions and slow processing times could result in missed learning opportunities, as noted by some students during the focus group discussions. Similar technical issues mentioned in previous studies disrupted the flow of scenarios.
and limited the effectiveness of the IVR learning experience [57,58]. These technical limitations must be addressed to ensure that IVR can be used effectively for teaching. Other main challenges that we experienced include the cost of implementing and maintaining the IVR technology, including hardware and software [50]. Another challenge is the need for technical support to develop and maintain IVR simulations, which requires collaboration between educators and technologists [59]. This may be prohibitive for some educational institutions to undertake.

Study Limitations

This study had several limitations that should be considered when interpreting the results. First, without a control group for comparison, it is unclear whether the positive outcomes identified from the surveys were based solely on this teaching innovation or because of the Hawthorn effect or the effect of novelty. However, the qualitative analyses were aligned with the survey findings, providing a more comprehensive understanding of this teaching innovation. Second, the use of the self-report method may have induced expectation bias. However, anonymity was adopted when conducting the questionnaires, which may have helped minimize bias. In addition, the objective evaluation of the students’ assignments strengthened the study by providing an independent measure of their attainment of the intended learning outcomes. Third, the students’ attainment of the learning outcomes was analyzed through a group assignment; thus, we could not differentiate between individual students in terms of performance. Fourth, the study population was restricted to one undergraduate nursing and OT cohort enrolled in a single university, thereby limiting the generalizability of the findings. Fifth, we were unable to confirm the reason behind the significant difference in empathy levels between nursing and OT students as it was beyond the scope of this study. Therefore, future studies are needed to explore the specific types of IVR teaching content suitable for enhancing empathetic feelings in undergraduate students from different health care professions. Sixth, we could not confirm the transferability of the knowledge obtained through IVR-assisted experiential learning to actual clinical practice.

Future Directions

To address the limitations of our study, we recommend conducting a randomized controlled trial with a control group in the future to evaluate the effects of IVR-assisted experiential tutorials on students’ empathy, learning experiences, and outcomes. In addition, individual assignments should be used to assess students’ attainment of the intended learning experiences and explore factors that could affect their performance. Such a study design would allow for a more robust evaluation of the effectiveness of IVR-assisted learning and provide deeper insights into the mechanisms underlying this approach. Moreover, future studies may be needed to determine whether the designs of related interventions have to be discipline specific to enhance empathy and understanding toward older adults with cognitive impairment among students of different health care disciplines. Further observational studies in clinical areas should also be considered to explore the transferability of knowledge to clinical practice regarding IVR-assisted experiential learning.

Conclusions

In conclusion, the findings of this study suggest that IVR-assisted experiential learning appears to have the potential to promote empathy and enhance the learning outcomes of undergraduate health care students regarding the care of older adults with cognitive impairment. Through immersive simulations, students were able to gain a deeper understanding of the challenges faced by this population and the importance of person-centered care. The findings also highlight the need to create age-friendly societies that reduce stigma, promote social inclusion, and provide community-based services and support. However, the challenges and limitations associated with the use of IVR for health care education must be addressed, such as technical issues, cost, and the need for technical support.

Acknowledgments

The authors sincerely thank the students and teachers who participated in this study. They also thank Ms. Jay Wong, the research assistant, for her excellent work in ensuring that this project ran smoothly. This study was funded by a Large-Scale Collaborative Teaching Development Grant (2019-2022) from the Learning and Teaching Committee and matching funds from the School of Nursing (.53.XX.49LP), Hong Kong Polytechnic University.

Data Availability

The data sets are not publicly available owing to pending further analysis of the data. The virtual reality scenarios used in this study were just a part of the scenarios present in the complete virtual reality training system. As we may need to conduct further comparisons, we regret not being able to disclose the data sets.

Authors’ Contributions

All the authors were involved in the design of the virtual reality (VR) games and the study. JYWL contributed to the conceptualization of the study. JYWL, DSKC, PPKK, KNKF, and TM implemented the VR games in their subjects. JYWL was responsible for data collection, data analysis, and quality control of the study. JYWL and PYM wrote the original draft of the manuscript. All coauthors commented on and rewrote the manuscript. All the authors have read and approved the final version of the manuscript.
Conflicts of Interest
None declared.

Multimedia Appendix 1
The findings on the between-group changes in Kiersma-Chen Empathy Scale scores across the time points between nursing and occupational therapy students.

[PDF File (Adobe PDF File), 13 KB - mededu_v10i1e48566_app1.pdf ]

References


43. Liu et al. JMIR MEDICAL EDUCATION


Abbreviations

IVR: immersive virtual reality
KCES: Kiersma-Chen Empathy Scale
OT: occupational therapy
PQ2: Presence Questionnaire version 2.0
VR: virtual reality

©Justina Yat Wa Liu, Pui Ying Mak, Kitty Chan, Daphne Sze Ki Cheung, Kin Cheung, Kenneth N K Fong, Patrick Pui Kin Kor, Timothy Kam Hung Lai, Tulio Maximo. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Measuring e-Professional Behavior of Doctors of Medicine and Dental Medicine on Social Networking Sites: Indexes Construction With Formative Indicators

Marko Marelić1, PhD; Ksenija Klasnić2, PhD; Tea Vukušić Rukavina1,3, MD, PhD

1Andrija Štampar School of Public Health, School of Medicine, University of Zagreb, Zagreb, Croatia
2Department of Sociology, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
3Biomedical Research Center Šalata, School of Medicine, University of Zagreb, Zagreb, Croatia

Corresponding Author:
Tea Vukušić Rukavina, MD, PhD
Andrija Štampar School of Public Health
School of Medicine
University of Zagreb
Rockfeller Street 4
Zagreb, 10000
Croatia
Phone: 385 14590126
Email: tvukusic@snz.hr

Abstract

Background: Previous studies have predominantly measured e-professionalism through perceptions or attitudes, yet there exists no validated measure specifically targeting the actual behaviors of health care professionals (HCPs) in this realm. This study addresses this gap by constructing a normative framework, drawing from 3 primary sources to define e-professional behavior across 6 domains. Four domains pertain to the dangers of social networking sites (SNSs), encompassing confidentiality, privacy, patient interaction, and equitable resource allocation. Meanwhile, 2 domains focus on the opportunities of SNSs, namely, the proactive dissemination of public health information and maintaining scientific integrity.

Objective: This study aims to develop and validate 2 new measures assessing the e-professional behavior of doctors of medicine (MDs) and doctors of dental medicine (DMDs), focusing on both the dangers and opportunities associated with SNSs.

Methods: The study used a purposive sample of MDs and DMDs in Croatia who were users of at least one SNS. Data collection took place in 2021 through an online survey. Validation of both indexes used a formative approach, which involved a 5-step methodology: content specification, indicators definition with instructions for item coding and index construction, indicators collinearity check using the variance inflation factor (VIF), external validity test using multiple indicators multiple causes (MIMIC) model, and external validity test by checking the relationships of the indexes with the scale of attitude toward SNSs using Pearson correlation coefficients.

Results: A total of 753 responses were included in the analysis. The first e-professionalism index, assessing the dangers associated with SNSs, comprises 14 items. During the indicators collinearity check, all indicators displayed acceptable VIF values below 2.5. The MIMIC model showed good fit ($\chi^2_{13}=9.4$, $P=.742$; $\chi^2$/df=0.723; root-mean-square error of approximation<.001; goodness-of-fit index=0.998; comparative fit index=1.000). The external validity of the index is supported by a statistically significant negative correlation with the scale measuring attitudes toward SNSs ($r=-0.225$, $P<.001$). Following the removal of 1 item, the second e-professionalism index, focusing on the opportunities associated with SNSs, comprises 5 items. During the indicators collinearity check, all indicators exhibited acceptable VIF values below 2.5. Additionally, the MIMIC model demonstrated a good fit ($\chi^2_4=2.5$, $P=.718$; $\chi^2$/df=0.637; root-mean-square error of approximation<.001; goodness-of-fit index=0.999; comparative fit index=1.000). The external validity of the index is supported by a statistically significant positive correlation with the scale of attitude toward SNSs ($r=0.338$, $P<.001$).

Conclusions: Following the validation process, the instrument designed for gauging the e-professional behavior of MDs and DMDs consists of 19 items, which contribute to the formation of 2 distinct indexes: the e-professionalism index, focusing on the dangers associated with SNSs, comprising 14 items, and the e-professionalism index, highlighting the opportunities offered by
SNSs, consisting of 5 items. These indexes serve as valid measures of the e-professional behavior of MDs and DMDs, with the potential for further refinement to encompass emerging forms of unprofessional behavior that may arise over time.

(JMIR Med Educ 2024;10:e50156) doi:10.2196/50156

**KEYWORDS**
e-professionalism; social media; formative index; social networking; doctors; medical; dental medicine

**Introduction**

**Background**
The development of social networking sites (SNSs) as a new form of media and communication channel has brought many changes to the health care system [1]. The widespread use of SNSs affects what we perceive as the professional behavior of health care professionals (HCPs) [2].

The rise in SNS users has sparked a growing interest in comprehending e-professionalism, particularly concerning SNSs. This specific facet of e-professionalism is becoming increasingly important. Over the past few years, numerous studies on the e-professionalism of HCPs have emerged [3,4], indicating a sustained momentum in generating scientific insights into e-professionalism.

**Defining and Measuring e-Professionalism**
The American Board of Internal Medicine (ABIM) guidelines on medical professionalism define 3 fundamental principles and a set of 10 professional responsibilities (or commitments). Fundamental principles are the importance of patient welfare, the principle of patient autonomy, and the principle of social justice. Professional responsibilities are commitments to professional competence, honesty with patients, patient confidentiality, maintaining appropriate relations with patients, improving the quality of care, improving access to care, a just distribution of finite resources, scientific knowledge, maintaining trust by managing conflicts of interest, and commitment to professional responsibilities [5].

E-professionalism is a specific type of professionalism. Cain and Romanelli [6] defined e-professionalism as the attitudes and behaviors (some of which may occur in private settings) reflecting traditional professionalism paradigms that are manifested through digital media.

A large number of previous research around e-professionalism measured the perception of e-professionalism [7-11] and attitude toward e-professionalism [12-18]. Through cross-validation, Kelley et al [19] created an instrument for measuring professional behaviors in pharmacy students, and even though there are some thematic overlaps, it is not suitable for measuring online behavior.

E-professionalism is often defined as a value which justifies the operationalization that directs the measurement of professionalism toward the measure of attitude. Nevertheless, from the perspective of the professions themselves, although professionalism is taught and transferred through socialization into the profession as a value, for assessing the level of e-professionalism of doctors of medicine (MDs) and doctors of dental medicine (DMDs) the behavioral component is of greater interest. Professional behavior, rather than just attitude, constitutes a visible aspect of professionalism. It is through professional behavior that not only patients and colleagues perceive a doctor’s professionalism, but also it is subject to internal control according to Freidsonian principles [20], enabling the profession to enforce sanctions on the professional. Professionalism is a behavior rather than an attitude because it should not be a hypothetical or idealized concept, as Evans [21] writes, but should be perceived as a reality—an actual entity. However, it is a real entity only if it is operational. To be real, professionalism must be something that people—professionals—actually “do,” not just something that the government or any other agency wants them to do, or wrongly imagines them to be doing [21]. The disconnection between behavior and attitude is termed “cognitive dissonance” [22], a phenomenon already acknowledged as a threat to the e-professionalism of HCPs on SNSs [4].

The research focused on the medical and dental professions as the target populations. These 2 fields were chosen due to their fundamental similarities, enabling comparisons, as well as their differences, suggesting potential variations in e-professionalism. Both medical and dental professions are sociologically recognized as professions [20] and share the commonality of providing health services. This entails a significant patient-practitioner relationship in both disciplines. Comparing various health professions is a valuable approach, and existing literature has already established overlaps in core competencies [23].

The primary distinction driving the selection of these 2 professions is the orientation of MDs, particularly in the Croatian context, toward the public sector, whereas DMDs are oriented toward the private sector.

This paper seeks to develop a reliable and valid instrument for assessing the e-professional behavior of both MDs and DMDs.

**Normative Framework for Defining e-Professional Behavior**

**Overview**
To define and measure e-professional behavior effectively, it is crucial to differentiate between professional and unprofessional behaviors. In our case, the primary objective of the normative framework is to delineate the content specifications, specifically the domains of instruments used to measure e-professional behavior.

The normative framework for assessing e-professionalism among MDs and DMDs draws upon 3 primary sources. While none of these sources alone is adequate for defining a comprehensive normative framework, each provides essential information crucial for its development. Some aspects of these
sources overlap conceptually, while others offer unique insights necessary for crafting the framework.

The first source comprises the e-professional conduct guidelines established by the ABIM [5]. These guidelines, among the earliest to be published, were developed through an international collaboration involving multiple institutions. They address the fundamental principles of professionalism and outline the professional responsibilities expected of MDs.

The second source consists of guidelines aimed at fostering e-professional behavior among medical and dental students [24]. While initially targeted at this specific demographic, a significant portion of the recommendations is applicable to the e-professionalism of MDs and DMDs. Consequently, these guidelines serve as a valuable resource for “reconstructing” the components of a normative framework for e-professionalism. They aid in delineating acceptable and unacceptable behaviors on SNSs within the context of medical and dental professions.

The third source is Julie Skrabal’s research [9], where she used the grounded theory method to develop a theoretical framework for e-professionalism. Her study empirically demonstrated which behaviors on SNSs are perceived as unprofessional. While the research focused on nursing students, the identification of key domains and indicators comprising professional behavior on SNSs holds significant value and applicability to MDs, DMDs, and all HCPs.

Based on the analysis of these 3 sources, e-professionalism, or e-professional behavior, can be categorized into 6 domains. Four of these domains pertain to the dangers associated with SNSs: confidentiality, privacy, contact with patients, and fair distribution of resources. The remaining 2 domains concern the opportunities afforded by SNSs: proactive dissemination of information relevant to public health and maintaining scientific objectivity. Each of these 6 domains is elaborated upon below.

**Confidentiality**

Confidentiality encompasses behaviors that primarily contravene the Health Insurance Portability and Accountability Act (HIPAA) of 1996. It entails safeguarding patient confidentiality to ensure that information regarding the patient is not disclosed, even to the patient’s relatives, without the patient’s explicit consent.

Concerning behavior on SNSs, HIPAA violations predominantly involve the unauthorized publication of photos or confidential patient information [9]. Additionally, adopting fake names (pseudonyms) to share posts containing medical or dental information constitutes another unprofessional behavior [24].

**Privacy**

This domain pertains to profile privacy settings and the management of post visibility. Barlow et al [25] established a correlation between privacy settings and unprofessional behavior, particularly among medical students. Consequently, they recommended the adoption of “private visibility settings” to mitigate such behaviors. Monitoring privacy settings [24], controlling post visibility [9,24], and seeking permission before tagging colleagues in posts to safeguard their privacy [24] are advocated practices. Furthermore, it is advisable to refrain from publishing professionally inappropriate content on SNSs, including posts containing curses, vulgar expressions, inappropriate attire, or behavior [9,24].

**Contact With Patients**

This domain encompasses direct contact with patients via SNSs. Inappropriate expressions, political incorrectness, or derogatory remarks toward patients or any individual or group can severely tarnish the public’s perception of doctors’ professional conduct [24]. Additionally, using unofficial channels, such as SNSs, to communicate sensitive professional information is considered unprofessional behavior within this domain [9].

**Fair Distribution of Resources**

Fair distribution of resources, as acknowledged in the ABIM guidelines, is considered an essential aspect of professional responsibility. While the ABIM guidelines emphasize the avoidance of unnecessary interventions and examinations, resource distribution also extends to SNSs. Time, a valuable resource allocated by MDs and DMDs to their patients, is particularly relevant in this context. Derived from the fundamental principle of professionalism known as the “Principle of Social Justice,” striving for a fair distribution of health care resources is imperative [5]. Communication with patients via SNSs typically requires the doctor’s time, often during their free time since it is an informal communication channel. According to the principle of fairness, it would be considered unprofessional behavior if a doctor selectively chooses which patients they are willing to communicate with on SNS and which they are not.

**Proactive Publication of Information of Public Health Interest**

The dimension of proactive publication of professional information of public health interest is one of the recognized aspects of e-professionalism that highlights the opportunity aspect of using SNSs. These behaviors are not deemed unprofessional when avoided; however, they can significantly contribute to e-professionalism when practiced by MDs and DMDs. While Skrabal [9] emphasizes creating positive postings as the absence of criticism and negative comments, proactive posting as a deliberate action toward e-professionalism is acknowledged in another research [26].

**Scientific Objectivity**

Sharing knowledge on SNSs is indeed desirable and constitutes professional behavior. However, it is essential to clearly differentiate between personal or subjective medical opinions and scientifically based facts [24].

**Formative Approach in Measuring e-Professionalism**

Most latent variables used in the social sciences are measured using reflective (effect) indicators [27,28]. According to a prevailing convention, indicators are seen as functions of the latent variable, whereby changes in the latent variable are reflected in changes in the observable indicators [27]. This is often true regarding constructs such as personality or attitude [28]. For example, attitude about SNSs affects respondents’ responses to the items posed to them. If someone has a negative attitude about SNSs, that attitude “guides” their responses.
However, in the case where the direction of “influence” is reversed, and where the indicators are “causing” the latent variable instead of “being caused by it,” then we can talk about formative measures [28].

Index construction focuses on explaining the abstract (unobserved) variance, considers multicollinearity among indicators, and emphasizes the role of an indicator as a “predictor” (latent variable) rather than “a predicted variable” [27].

The choice of approach (reflexive vs formative) stems from the concept, that is, from the relationship between variables and constructs [29]. Jarvis et al [30] stated 4 conditions that can help discern whether a reflective or a formative model is appropriate: (1) the direction of causality between the construct and the indicator, (2) the interchangeability of the indicators, (3) covariance between indicators, and (4) the nomological network of construct indicators.

The first argument presented by Jarvis et al [30] is valid for our research because, unlike attitude, e-professional behavior stems from specific actions and decisions on SNSs. If someone refrains from posting pictures of patients, seeks permission from a colleague before mentioning them on SNS, actively controls the visibility of their posts, and takes similar actions, then these decisions contribute to their e-professional behavior.

For the second argument, e-professional behavior indicators are not interchangeable, even though they all measure e-professionalism. Posting a picture of a patient on an SNS is considered unprofessional behavior, but so is posting pictures from parties at work. Both behaviors are unprofessional, although they are not interchangeable in measurement (someone may frequently post photos of patients but rarely post workplace-related images).

The third argument states that covariance among indicators is unnecessary [30]. It is neither expected nor needed here because recognized behaviors within the normative framework can be entirely unrelated but still measure e-professional behavior (e.g., sending a friend request to a patient and asking a colleague to mention them in a post).

The fourth argument suggests that the nomological network in the formative model can have different antecedents and consequences [30]. Indicators of e-professional behavior do not need to share the same antecedents because they can be driven by different motivations. A doctor may post pictures of patients because they believe it raises awareness about a particular illness (even though this act is unprofessional), while the motivation for unprofessional behavior, such as posting pictures from workplace parties, does not stem from the same motivation.

Based on these arguments, the behavioral component of e-professionalism measured in this paper conceptually corresponds to the formative approach.

We presume that other research in this area has not applied a formative approach in measuring e-professionalism because they have yet to define e-professionalism as a behavior.

Diamantopoulos and Winklhofer [28] proposed 4 key steps for validating indexes with formative indicators. The first step, content specification, refers to specifying the scope of the latent variable; in the second step, it is necessary to define the indicators; the third step refers to checking the collinearity of the indicators using the variance inflation factor (VIF) [28]. The fourth step is to assess the external validity of the index. Verification of the external validity of formative indices is often carried out by checking the relationship of the index with other measures and variables (as cited in [28]).

Although these 4 steps are sufficient for constructing and validating the index, it is possible to make an additional check of the external validity proposed by Diamantopoulos and Winklhofer [28]. This requires creating a model in which some reflective indicators are included (Diamantopoulos and Winklhofer [28] use 2) in the same model as the formative indicators. This model is called the multiple indicators multiple causes (MIMIC) model [28]. Acceptable overall model fit suggests retention of items in the formative model. If the exclusion of some items can significantly increase the model fit under the very strict condition that not a single exclusion would violate the content validity of the formative model, only then can the items be excluded.

In this paper, we have followed these 4 key steps for validating indexes with formative indicators. An additional step (the MIMIC model) was conducted before assessing the external validity of the index.

Methods

Sample

Quantitative survey data were collected using an online survey questionnaire. The Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [31] is available in Multimedia Appendix 1. The required sample size was defined according to a conservative estimate often used for multivariate analyses, corresponding to a 10:1 ratio between the number of observations and the number of variables used in the questionnaire’s largest instrument [32]. In our case, that is a sample size of 280 (140 MDs and 140 DMDs). The type of sample was a nonprobabilistic purposive sample.

The study was a part of a long-term research project funded by the Croatian Science Foundation, UIP-05-2017 “Dangers and Benefits of Social Networks: E-Professionalism of Health Care Professionals – SMePROF” [33].

The mailing lists used to distribute the survey were the official full membership emailing lists of the Croatian Medical Chamber (CMC) and Croatian Chamber of Dental Medicine (CCDM). At the time of the survey, the CMC’s emailing list contained 15,562 email addresses of MDs, and the CCDM’s emailing list contained 7616 email addresses of DMDs. The email included a brief text about the study’s objective, the expected time to complete the survey, and the person and university responsible for conducting the study.

Participation in the survey was voluntary; there was no form of incentive to complete the survey. To ensure anonymity, no identification data were collected. Data were collected from February to July 2021, with 2 reminders sent in that period.
The MIMIC model was used to check the external validity of multicollinearity of the instruments was tested with the VIF. The coefficient and phi coefficients of associations. The quantitative variables were tested with the Pearson correlation determined depending on appropriateness. Correlations between mean, range, SDs, and responses was carried out, and distribution measures such as A descriptive analysis of frequencies and percentage of

Measures

The instrument for measuring the e-professional behavior of MDs and DMDs, presented in this study, is part of a more extensive questionnaire called SMePROF Project Survey Questionnaire on Social Media Usage, Attitudes, Ethical Values and E-professional Behaviour of Doctors of Medicine and Doctors of Dental Medicine, available at Viskić et al [34]. Although the questionnaire contained multiple instruments partially derived from previous studies [10,34,35], the instrument for measuring the e-professional behavior of MDs and DMDs is a novel instrument created by the authors. The instrument contains 20 items measured using the self-reporting approach, used to create 2 e-professionalism indexes, and the process is explained in the following parts of this paper. In validating indexes, an MIMIC model was used, which required 4 reflexive variables $\gamma_1$-$\gamma_4$ measuring attitude toward e-professionalism. These items were taken from a validated instrument for measuring attitudes toward e-professionalism [35]. Descriptive of these 4 reflexive variables are shown in Multimedia Appendix 2. The MIMIC model was exclusively used as a method for validating the external validity of the indexes, and not for theory development.

The associations of indexes with theoretically related constructs were tested to assess the external validity. For this purpose, we used a validated instrument for measuring attitudes toward SNSs [36]. The instrument was translated into the Croatian language, and after additional reliability checks, 1 item was removed from the scale (“Potential and/or existing employers may use the information found on SNS to make decisions about prospective and/or existing employees”). The final instrument used had 12 items and Cronbach $\alpha=.70$.

Analytical Methods

A descriptive analysis of frequencies and percentage of responses was carried out, and distribution measures such as mean, range, SDs, and $\alpha_3$ measure of asymmetry were determined depending on appropriateness. Correlations between quantitative variables were tested with the Pearson correlation coefficient and phi coefficients of associations. The multicollinearity of the instruments was tested with the VIF. The MIMIC model was used to check the external validity of instruments with formative indicators. Data analysis was performed using IBM SPSS Statistics 26. IBM SPSS Amos 22 was used to test the MIMIC model.

Results

Survey Responses

A total of 1013 responses were collected. The response rate was 4.37% (1013/23,178). The final realized sample of the entire research contained the answers of 999 respondents, of which 75.4% (753/999) use at least one SNS, 67.3% (507/753) of the respondents were MDs and 32.7% (246/753) were DMDs. The sample was predominantly female (558/753, 74.1%) with an average age of 38 (SD 10.99) years. Most respondents worked in a public health institution (412/753, 54.7%), and the second most frequent type of workplace was a private institution with a contract with the Croatian Health Insurance Fund (CHIF; 148/753, 19.7%).

Previous research on the same sample [34] showed a significant difference in age, where MDs were older than DMDs with an average age of 39.26 years as opposed to 36.58 years, respectively, and in the type of employment, with more than two-thirds of DMDs (168/246, 68.2%) being employed in the private sector compared with only 20.5% (104/507) of MDs.

All specialization status levels are included in the sample (Multimedia Appendix 3).

The Construction of the e-Professionalism Index—The Danger Aspect of SNSs

Following the first step in creating the index, according to Diamantopoulos and Winklhofer [28], the content for the latent variable is specified below. In the second step, e-professional behaviors described in the normative framework were operationalized into an instrument for measuring the aspect of e-professionalism related to the dangers of SNSs (Table 1). The identified indicators are grouped into 4 domains: confidentiality, privacy, contact with patients, and fair distribution of resources. Items were evaluated on a frequency rating scale: 0=Never, 1=Rarely, 2=Occasionally, 3=Often; and the option “I have never been in a situation where this could happen” was added. It was essential to distinguish behaviors that could have happened but did not from those for which the respondent was not even in a situation to practice them. Depending on the direction and content of the items, the difference between the opportunity to behave in a certain way and the frequency of that behavior can mean the difference between professional and unprofessional behavior. In the case of items formulated in a positive direction (marked +), a higher frequency measures a higher level of e-professionalism. In the case of items formulated in a negative direction (marked –), higher frequency measures a lower level of e-professionalism.
Table 1. Domains, indicators, and items for the instrument of e-professionalism—the danger aspect of SNSs.

<table>
<thead>
<tr>
<th>Domain and indicator</th>
<th>Item</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Confidentiality</strong></td>
<td>Disclosure of patient information.</td>
<td>I published some information about my patient.</td>
</tr>
<tr>
<td></td>
<td>Publication of photographs of the patient without their consent.</td>
<td>I posted a photo of my patient without their knowledge.</td>
</tr>
<tr>
<td></td>
<td>Hiding behind false names when posting online or anonymously posting medical information.</td>
<td>I shared medical/dental advice on SNS without my name being visible.</td>
</tr>
<tr>
<td></td>
<td>Confidentiality of communication also applies to SNS.</td>
<td>I shared some information about the patient I received through SNS with others.</td>
</tr>
<tr>
<td><strong>Privacy of MDs and DMDs profiles</strong></td>
<td>Active management of the visibility of posts depending on their content.</td>
<td>Depending on the appropriateness of the content of my posts, I determine to whom they will be visible.</td>
</tr>
<tr>
<td></td>
<td>Controlling the visibility of other people’s posts that include you, depending on their content.</td>
<td>If I notice that someone else has published something about me (eg, my picture, location, or similar), I control who will see it.</td>
</tr>
<tr>
<td></td>
<td>Seeking prior approval from colleagues to publish information about them.</td>
<td>I asked a colleague’s permission to mention them in the post.</td>
</tr>
<tr>
<td></td>
<td>Appropriate behavior on published content from a professional context.</td>
<td>I have posted content that shows informal situations at my workplace (eg, drinks with colleagues or parties at work).</td>
</tr>
<tr>
<td></td>
<td>The use of profanity and other vulgar expressions in posts.</td>
<td>A curse word or some different vulgar expression occasionally slips out in my posts.</td>
</tr>
<tr>
<td><strong>Contact with patients</strong></td>
<td>Inappropriate expression in posts.</td>
<td>In my posts, I am cautious that my expression is entirely professional.</td>
</tr>
<tr>
<td></td>
<td>Separation of professional and private communication.</td>
<td>I communicate with patients regarding medical/dental problems and treatment from a private profile.</td>
</tr>
<tr>
<td></td>
<td>Inclusion of patient data obtained at SNS in the medical documentation without the patient’s consent.</td>
<td>I included information about the patient I found through SNS in the medical documentation without their knowledge.</td>
</tr>
<tr>
<td></td>
<td>Sending a friend request to a patient or a patient’s family member.</td>
<td>Have you ever sent a “friend request” to a patient or a member of the patient’s family from a private profile on an SNS?</td>
</tr>
<tr>
<td><strong>Fair distribution of resources</strong></td>
<td>Communication with patients via SNS and outside working hours is selective (the doctor chooses whom they respond to; patients without SNS cannot reach them).</td>
<td>On SNS, I choose which patients I will make contact with and which I will not.</td>
</tr>
</tbody>
</table>

*SNS: social networking site.

bFor items formulated in a positive direction (marked +), a higher frequency measures a higher level of e-professionalism. In the case of items formulated in a negative direction (marked –), a higher frequency measures a lower level of e-professionalism.

The indicator “Sending a friend request to a patient or a member of the patient’s family” was not measured as frequency. Instead, the 4 offered answers were as follows: Yes, to the patient; Yes, to a family member; Yes, both; and No. The negative response is considered professional, while all other responses indicate unprofessional behavior.

The descriptive results for the items that measure the aspect of e-professionalism related to the dangers of SNSs are shown in Table 2. The items that measure e-professional behavior are marked with a “b.” All other items measure e-unprofessional behavior.
Table 2. E-professionalism (the dangers aspect of SNSs\textsuperscript{a}) descriptives (N=753).

<table>
<thead>
<tr>
<th>Danger aspects</th>
<th>Never, n (%)</th>
<th>Rarely, n (%)</th>
<th>Occasionally, n (%)</th>
<th>Often, n (%)</th>
<th>I have never been in a situation where this could happen, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I asked a colleague’s permission to mention them in the post.\textsuperscript{b}</td>
<td>170 (22.6)</td>
<td>117 (15.5)</td>
<td>71 (9.4)</td>
<td>50 (6.6)</td>
<td>345 (45.8)</td>
</tr>
<tr>
<td>2. I shared some information about the patient that I received through SNS with other people.</td>
<td>368 (48.9)\textsuperscript{c}</td>
<td>61 (8.1)</td>
<td>26 (3.5)</td>
<td>3 (0.4)</td>
<td>295 (39.2)</td>
</tr>
<tr>
<td>3. I posted a photo of my patient without their knowledge.</td>
<td>492 (65.3)\textsuperscript{c}</td>
<td>14 (1.9)</td>
<td>7 (0.9)</td>
<td>2 (0.3)</td>
<td>238 (31.6)</td>
</tr>
<tr>
<td>4. I included information about the patient I found through SNS in the medical documentation without their knowledge.</td>
<td>484 (64.3)\textsuperscript{c}</td>
<td>3 (0.4)</td>
<td>2 (0.3)</td>
<td>0 (0.0)</td>
<td>264 (35.1)</td>
</tr>
<tr>
<td>5. I shared medical/dental advice on SNS without my name being visible.</td>
<td>503 (66.8)\textsuperscript{c}</td>
<td>39 (5.2)</td>
<td>6 (0.8)</td>
<td>4 (0.5)</td>
<td>201 (26.7)</td>
</tr>
<tr>
<td>6. Depending on the appropriateness of the content of my posts, I determine to whom they will be visible.\textsuperscript{b}</td>
<td>295 (39.2)</td>
<td>99 (13.1)</td>
<td>93 (12.4)</td>
<td>71 (9.4)\textsuperscript{c}</td>
<td>195 (25.9)</td>
</tr>
<tr>
<td>7. If I notice that someone else has published something about me (eg, my picture, location, or similar), I control who will see it.\textsuperscript{b}</td>
<td>209 (27.8)</td>
<td>106 (14.1)</td>
<td>104 (13.8)</td>
<td>177 (23.5)\textsuperscript{c}</td>
<td>157 (20.8)</td>
</tr>
<tr>
<td>8. I have published content that shows informal situations at my workplace (eg, drinks with colleagues or parties at work).</td>
<td>354 (47.0)\textsuperscript{c}</td>
<td>181 (24.0)</td>
<td>84 (11.2)</td>
<td>17 (2.3)</td>
<td>114 (15.1)</td>
</tr>
<tr>
<td>9. I published some information about my patient.</td>
<td>579 (76.9)\textsuperscript{c}</td>
<td>22 (2.9)</td>
<td>5 (0.7)</td>
<td>2 (0.3)</td>
<td>145 (19.3)</td>
</tr>
<tr>
<td>10. I communicate with patients regarding medical/dental problems and treatment from a private profile.</td>
<td>423 (56.2)\textsuperscript{c}</td>
<td>133 (17.7)</td>
<td>64 (8.5)</td>
<td>14 (1.9)</td>
<td>119 (15.8)</td>
</tr>
<tr>
<td>11. On SNS, I choose which patients I will make contact with and which I will not.</td>
<td>293 (38.9)\textsuperscript{c}</td>
<td>74 (9.8)</td>
<td>65 (8.6)</td>
<td>76 (10.1)</td>
<td>245 (32.5)</td>
</tr>
<tr>
<td>12. In my posts, I am cautious that my expression is entirely professional.\textsuperscript{b}</td>
<td>51 (6.8)</td>
<td>61 (8.1)</td>
<td>111 (14.7)</td>
<td>366 (48.6)\textsuperscript{c}</td>
<td>164 (21.8)</td>
</tr>
<tr>
<td>13. A curse word or some other vulgar expression occasionally slips out in my posts.</td>
<td>494 (65.6)\textsuperscript{c}</td>
<td>86 (11.4)</td>
<td>23 (3.1)</td>
<td>2 (0.3)</td>
<td>148 (19.7)</td>
</tr>
<tr>
<td>14. Have you ever sent a “friend request” to a patient or a member of the patient’s family from a private profile on an SNS\textsuperscript{d}\textsuperscript{c}</td>
<td>699 (92.8)\textsuperscript{c}</td>
<td>33 (4.4)</td>
<td>3 (0.4)</td>
<td>18 (2.4)</td>
<td>N/A\textsuperscript{e}</td>
</tr>
</tbody>
</table>

\textsuperscript{a}SNS: social networking site.  
\textsuperscript{b}Item represents professional behavior on SNS.  
\textsuperscript{c}Response represents professional behavior on SNS.  
\textsuperscript{d}The options were “no,” “yes, to a patient,” “yes, to a family member,” and “yes, both,” respectively.  
\textsuperscript{e}N/A: not applicable.  

The answer “I have never been in a situation where this could happen” is not a missing value, but it carries a conceptual meaning that must be distinguished from the answer “Never.” The assessment of whether that answer is professional or unprofessional depends on the content and direction of the item. Respondents who have never engaged in unprofessional behavior are professional, but so are those who never had an opportunity to act unprofessionally. Respondents who often practice behaviors on items marked with “b” are professional, and so are those who have never been in a situation to practice these behaviors because they have not been in a situation to behave unprofessionally. For example, in the case of positive items (those representing professional behavior), such as “I asked a colleague’s permission to mention him/her in the post,” professional behavior is defined as a situation where the individual has never violated this rule because they have never mentioned colleagues in their posts or seek permission each time they mention them. Any other frequency level implies that, at some point, the person has posted about colleagues without their consent, which constitutes unprofessional behavior on SNS.
It is crucial here to differentiate between the absence of behavior of interest (requesting permission from colleagues when mentioning them in posts) in situations where it should have been sought (if mentioning them in posts) from the situations where it should not have been sought (because they never mention colleagues).

By contrast, for negative items (those representing unprofessional behavior), such as “I shared some information about the patient that I received through SNS with other people,” professional behavior is defined as situations where the individual has never engaged in such behavior or has not even been in a situation where they could engage in such behavior (eg, they do not communicate with patients via SNS, so they cannot receive patient information through this channel).

Therefore, the context of the absence of specific behaviors plays a pivotal role in distinguishing between professional and unprofessional behaviors. It is essential to combine the response “I have never been in a situation where this could happen” with the level of behavior frequency.

To construct the index, the frequency of behavior on each indicator was not graded but only considered as a binary value (professional vs unprofessional).

For items that measure unprofessional behavior, any degree of frequency other than “never” was considered unprofessional behavior. For items that measure professional behavior (eg, asking a colleague’s permission to mention them in a post), all those who did this never, rarely, or occasionally were considered unprofessional on that indicator, because this is the behavior they are expected to do always (or often in our scale).

The Validation Process of the e-Professionalism Index—The Danger Aspect of SNSs

After specifying the scope and defining the indicators, the third step, according to Diamantopoulos and Winklhofer [28], refers to checking the collinearity of the indicators. Intercorrelations of the items in the e-professionalism instrument—the danger aspect of SNSs are shown in Multimedia Appendix 4. Given that these are binary variables, phi coefficients of associations were used. The correlation between the variables “On SNS, I choose with which patients I will make contact with and which I will not” and “From a private profile, I communicate with patients regarding medical/dental problems and treatment” \((r=0.568)\) represents a moderate correlation and evokes the need to investigate potential multicollinearity. This suggests that those who communicated with patients via SNSs also chose with whom (patients) they would establish communication. As a formative approach is used, special care is needed before excluding indicators to preserve the instrument’s validity. Therefore, the VIF and MIMIC model were calculated. Multicollinearity was tested using a VIF with an additive index of e-professionalism, an aspect of the danger of SNS that was constructed as the sum of values on binary indicators. According to the conservative threshold [37], VIF values on all indicators were below the value of 2.5, which suggests that multicollinearity is not an issue.

The MIMIC model was implemented to check the external validity of the instrument. The path diagram of the MIMIC model is shown in Multimedia Appendix 5. Variables \(x_1-x_{14}\) correspond to the items from Table 2. Items \(y_1\) (Communication with a patient through social media can be achieved without compromising doctor-patient confidentiality) and \(y_2\) (Social media have the potential to improve communication between a doctor and a patient) were chosen as reflective indicators.

The model showed good fit \((\chi^2_{13}=9.4, P=.742; \chi^2/df = .723; \text{root-mean-square error of approximation}<0.001; \text{goodness-of-fit index}=0.998; \text{comparative fit index}=1.000). However, 7 of the 14 items \((x_1, x_2, x_3, x_6, x_7, x_8, \text{and } x_{13})\) did not have significant regression coefficients \(\gamma\) that can also be interpreted as validity coefficients [28]. The probable reason is that the measured reflective indicators did not measure the same domains as e-professional behavior; instead, they measured an attitude toward e-professionalism. Both items 11 \((P<.001)\) and 12 \((P=.02)\), which were investigated as potential problems of multicollinearity, have significant validity coefficients. Considering that, as well as an acceptable VIF, they were retained in the index to preserve the content validity to which formative models are particularly sensitive.

A higher value on the index means a higher degree of e-professionalism, that is, a lower incidence of unprofessional behavior on SNSs. The index results ranged from 0 to 14, and the average value in our sample was 10.60 (SD 2.173). The distribution of the index was skewed toward higher values \((\alpha=-.44, P=.09)\), that is, toward the professional behavior of our respondents on SNSs.

The external validity of the index is supported by the correlation with other measured constructs. There was a statistically significant negative correlation between the index of e-professionalism (aspects of the danger of SNSs) and the scale of attitude toward SNSs \((r=-.225, P<.001)\).

The Construction of the e-Professionalism Index—The Opportunity Aspect of SNSs

The construction of the e-professionalism index—the opportunity aspect of SNSs follows the same validation steps as the aspect of the dangers of SNSs [28].

E-professional behaviors described in the normative framework were operationalized into an instrument for measuring e-professionalism through the opportunity aspect of SNSs. The instrument contains 2 domains, measured by 6 items. All items are formulated in the same direction so a higher frequency measures a higher level of e-professionalism (Table 3).
Table 3. Domains, indicators, and items for the instrument of e-professionalism—opportunity aspect of SNSs.\(^a\)

<table>
<thead>
<tr>
<th>Domain and indicator</th>
<th>Item</th>
<th>Direction(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proactive posting of expert information of public health interest</td>
<td>I share posts on social media that contain general medical advice.</td>
<td>+</td>
</tr>
<tr>
<td>Sharing posts that contain general medical advice</td>
<td>I share posts on social media that contain general medical advice.</td>
<td>+</td>
</tr>
<tr>
<td>Sharing new scientific knowledge in the field of medicine</td>
<td>I use my profile to share information about new scientific knowledge in the field of medicine.</td>
<td>+</td>
</tr>
<tr>
<td>Debunking medical myths and misinformation</td>
<td>I debunk medical/dental myths and misinformation by posting on SNS.</td>
<td>+</td>
</tr>
<tr>
<td>Calling for public health actions</td>
<td>I use SNS to raise public awareness of public health actions.</td>
<td>+</td>
</tr>
<tr>
<td>Encouraging responsible behavior</td>
<td>I create posts on SNS that call for responsible health behavior.</td>
<td>+</td>
</tr>
</tbody>
</table>

Scientific objectivity

Emphasis on distinguishing personal medical opinions from facts

In the posts, I clearly separate my personal opinion on a medical/dental issue from scientifically confirmed facts.

\(^a\)SNS: social networking site.

\(^b\)All items are formulated in the same direction so a higher frequency measures a higher level of e-professionalism.

The descriptive results for the items that measure the opportunity aspect of SNSs are shown in Table 4. While measuring the danger aspect of SNSs focused on occurrence, not on the frequency of occurrence, the frequency of each behavior is relevant with this instrument. All behaviors in this instrument have the characteristic of being desirable, but the absence of such behaviors is not unprofessional. If an MD or DMD practices these behaviors, they use opportunities of SNSs and contribute to their professionalism. However, if they do not practice any of these behaviors, or have never been in a situation where they can behave like that, it is not unprofessional, but misses the opportunity to use the advantages of SNSs.

Table 4. E-professionalism—opportunity aspect of SNSs.\(^a\)—descriptives (N=753).

<table>
<thead>
<tr>
<th>Item</th>
<th>Never, n (%)</th>
<th>Rarely, n (%)</th>
<th>Occasionally, n (%)</th>
<th>Often, n (%)</th>
<th>I have never been in a situation where this could happen, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I debunk medical/dental myths and misinformation by posting on SNS.</td>
<td>355 (47.1)</td>
<td>130 (17.3)</td>
<td>87 (11.6)</td>
<td>16 (2.1)</td>
<td>165 (21.9)</td>
</tr>
<tr>
<td>2. I share posts on social media that contain general medical/dental advice.</td>
<td>312 (41.4)</td>
<td>167 (22.2)</td>
<td>128 (17.0)</td>
<td>28 (3.7)</td>
<td>118 (15.7)</td>
</tr>
<tr>
<td>3. I use SNS to raise public awareness of public health actions.</td>
<td>185 (24.6)</td>
<td>191 (25.4)</td>
<td>224 (29.7)</td>
<td>84 (11.2)</td>
<td>69 (9.2)</td>
</tr>
<tr>
<td>4. I use my profile to share information about new scientific knowledge in the field of medicine/dental medicine.</td>
<td>248 (32.9)</td>
<td>183 (24.3)</td>
<td>184 (24.4)</td>
<td>54 (7.2)</td>
<td>84 (11.2)</td>
</tr>
<tr>
<td>5. I create posts on SNS that call for responsible health behavior.</td>
<td>186 (24.7)</td>
<td>196 (26.0)</td>
<td>212 (28.2)</td>
<td>81 (10.8)</td>
<td>78 (10.4)</td>
</tr>
<tr>
<td>6. In the posts, I clearly separate my personal opinion on a medical/dental issue from scientifically confirmed facts.</td>
<td>170 (22.6)</td>
<td>49 (6.5)</td>
<td>77 (10.2)</td>
<td>149 (19.8)</td>
<td>371 (49.3)</td>
</tr>
</tbody>
</table>

\(^a\)SNS: social networking site.

To construct an index reflecting the degree of e-professionalism in utilizing social networking opportunities, responses marked as “Never” or “I have never been in a situation where this could happen” are not considered contributions to e-professionalism and are coded as 0. Conversely, responses categorized as “Rarely,” “Occasionally,” and “Often” contribute to e-professionalism, representing 3 levels of engagement with the benefits of social networks and are coded as 1, 2, and 3 respectively.

The Validation Process of the e-Professionalism Index—Opportunity Aspect of SNSs

The correlations between the items that constitute this index have higher values than those in the aspect of dangers of the SNS index (Multimedia Appendix 6). The item “I create posts on SNS that call for responsible health behavior” moderately correlates with several items (from \(r=0.418\) to 0.714). To check
if multicollinearity is present in this instrument, paying attention to the VIF is necessary.

VIF was calculated with an additive index of e-professionalism—opportunity aspect of SNSs. VIF values on all indicators are below the value of 2.5, which suggests no risk of multicollinearity, even according to a conservative interpretation.

Before excluding the item “I create posts on SNS that call for responsible health behavior,” an MIMIC model was created with all the items included, and a second model without that item was created to check for any changes in the model fit. The diagram of the MIMIC model is shown in Multimedia Appendix 5. Variables $x_1$-$x_8$ correspond to the items from Table 4. Items $y_1$ (As MD/DMD, it is my duty to keep abreast of current trends in the use of SNS) and $y_2$ (Guiding patients to online information is a new responsibility of MDs/DMDs in the digital age) were chosen as 2 reflective indicators.

The MIMIC model with all 6 items showed good fit characteristics ($\chi^2=2.880$, $P=.718$; $\chi^2/df = 0.576$; root-mean-square error of approximation=.001; goodness-of-fit index=.999; comparative fit index=1.000). However, 3 items ($x_1$, $x_3$, and $x_8$) did not have significant regression coefficients ($\gamma$, $P=.14$, $P=.44$, and $P=.19$, respectively).

Considering the high correlations with other items, the VIF value that exceeds the limit of 2.5, and the regression coefficient $\gamma$ that is not statistically significant ($P=.19$), item $x_5$ was excluded from the e-professionalism index—opportunity aspect of SNSs. After excluding item $x_5$, the fit of the MIMIC model did not change significantly ($\Delta \chi^2=0.336$, $P=.56$) and the fit of the model was $\chi^2=2.544$, $P=.718$; $\chi^2/df = 0.637$; root-mean-square error of approximation=.001; goodness-of-fit index=.999; comparative fit index=1.000.

The index of e-professionalism—opportunity aspect of SNSs was created as the sum of the values of the remaining 5 recoded variables. A higher value on the e-professionalism index means more inclined they are to use them when working with patients, the more positive attitude the respondents have about SNSs, the more likely they will take advantage of the benefits of SNSs.

In the index of e-professionalism—the danger aspect of SNSs, all initially operationalized indicators were retained. In the index of e-professionalism—the opportunity aspect of SNSs, item $x_5$ (I create posts on SNS that call for responsible health behavior) measuring the indicator “Encouraging responsible behavior” was excluded. The formative approach suggests cautious consideration of managing the content validity of the model. It seems that respondents understood item $x_5$ very similarly to item $x_3$ (I use SNS to raise public awareness of public health actions.). After testing the indicators in the MIMIC model, the authors concurred that the content validity is not threatened by excluding this item, and multicollinearity would pose a more significant problem than losing a very subtle difference in the contents of these items.

### Comparison With Prior Work

Conceptual domains recognized in this study only partially overlap with domains in the instrument of (offline) professional behavior [19] and the instrument for measuring attitudes toward e-professionalism [35]. Kelley et al [19] recognized a domain called “Upholding principles of integrity and respect,” which corresponds to the domain “Confidentiality” in this study, as well as “Citizenship and professional engagement” [19], which corresponds to “Proactive posting of expert information of public health interest.” In an instrument for measuring attitudes toward e-professionalism, Marelić et al [35] recognized the domain “Ethical aspects” that theoretically includes HIPAA violations and therefore corresponds to the domain “Confidentiality” in this study, and the domain “Physicians
the digital age” that corresponds to “Contact with patients”. However, the instrument of (offline) professional behavior contains domains that are not comparable to e-professional behavior, and the instrument for measuring attitudes toward e-professionalism contains domains that are not applicable for behavior measurement, and because of potential cognitive dissonance, measuring attitude is not a replacement for behavior measurement.

Limitations
The first limitation of this study is the low response rate (1013/23,178, 4.37%). Previous research has indicated that these professions have low survey response rates, especially in e-mailing surveys using web-based formats [38-42]. Time, confidentiality concerns, and topic relevance are some of the main reasons for their low survey participation [40]. Previous research has indicated that declining response rates among HCPs may be attributed to various factors, including heightened requests to participate in surveys and increased workloads. This increase in workload encompasses both the rising number of patients and administrative responsibilities [38,39].

One factor likely contributing to the low response rate in this study is the demanding schedule of MDs and DMDs. The estimated time required to complete our survey was lengthy, ranging from 10 to 15 minutes, due to the inclusion of a complex and comprehensive questionnaire containing 40 questions. Moreover, the survey was conducted during the COVID-19 pandemic (February to July 2021), a period marked by heightened strain on the health care system. MDs, especially those in Croatia, were confronted with extreme workloads and specific working conditions during this time. Additionally, MDs received numerous invitations to participate in web-based surveys, particularly regarding the impact of the COVID-19 pandemic on their physical or mental health. Given these circumstances, our study’s focus on e-professionalism may have been perceived as of lower interest, potentially further reducing doctors’ willingness to participate in research.

However, our objective in creating and validating new indexes did not prioritize achieving representativeness in our sample or generalizing our findings to the entire population of MDs and DMDs in Croatia. Instead, our focus was on assessing the suitability of the developed measurement instruments across various medical professions, using a nonprobabilistic purposive sample. Our final sample comprised responses obtained from the population of interest for this study, specifically MDs and DMDs who use at least one SNS. It is worth noting that the number of responses received in our survey (507 MDs and 246 DMDs) exceeded the initially planned sample size (140 MDs and 140 DMDs) by a considerable margin.

The second limitation concerns a relatively large proportion of respondents (ranging from 69/753, 9.2%, to 371/753, 49.3%) who selected the option “I have never been in a situation where this could happen” for certain items. It remains unclear why they did not simply respond with “Never.” The reasons behind this choice are ambiguous. It is possible that some respondents are passive users of SNSs, thus not engaging in any content publication and consequently unable to exhibit unprofessional behavior. Alternatively, it could be that these respondents do not work directly with patients, rendering items related to violations of the HIPAA irrelevant to them. Another possibility is that they perceive their standards of professionalism to be exceptionally high, leading them to believe they would never engage in such behavior. While this issue does not affect the measurement of the occurrence of e-(un)professional behavior, it does impede a detailed understanding of the frequency of e-unprofessional behavior. Addressing this limitation could be a focus of future research and modifications to the measurement instrument, but this should be preceded by gaining new insights into the e-professional behaviors of MDs and DMDs.

The third limitation involves the potential for bias associated with using a self-reporting approach to measurement. Similar to other self-report measures in medicine, 2 key biases often arise: recall bias and social desirability bias [43]. Recall bias in our study could be attributed to the lack of a specified timeframe, such as “during the last year.” We chose this approach because it represents the initial assessment of such behaviors, and we faced a scarcity of existing data on this subject. Introducing a specific timeframe in future research could aid in mitigating potential recall bias. The potential for social desirability bias stems from 2 sources. First, the nature of the measurement itself requires HCP respondents to self-report potentially unprofessional behaviors, including some that may constitute violations of HIPAA. The other factor to consider is that respondents were contacted to participate in our research through the same institutions responsible for granting and revoking licenses to practice medicine/dental medicine. Despite our assurance of anonymity in the study, respondents may have felt compelled to provide socially desirable answers on certain items. One method to mitigate or control social desirability bias is to include positive items, such as those measuring professional behaviors, alongside other items. An additional approach to address both biases, which could serve as a recommendation for future research, involves further refinement and validation of the instrument. This could be achieved by comparing self-reported data with information obtained through web scraping of respondents’ SNS profiles, particularly focusing on visible behaviors.

The fourth limitation arises from the potential mismatch between the use of reflective indicators \( y_1, y_2 \) in the MIMIC model and the nature of the created indexes, which are intended to measure e-professionalism as behavior. However, the reflective variables used in the model measure attitude. While this approach was necessary for creating the MIMIC model in this study, there is a possibility that cognitive dissonance [4,21] may compromise the fit of the model.

The fifth limitation to note is that the sources used to establish a normative framework were relevant to the time and location of this research. However, their applicability to other countries and populations of HCPs, or their accuracy over time, may be limited. For example, the ABIM e-professional conduct guidelines [5] are relatively dated, and while they represent fundamental values of professionalism, they may not fully encompass changes in societal values that have occurred since the emergence of SNSs. Specific behaviors measured in these indexes may require revision or supplementation in the future.
Moreover, additional studies conducted after the development of this index may offer new insights into creating a normative framework for defining e-professional behaviors [44].

Future Directions
In considering avenues for enhancing both the instruments used in this study and future research directions, it becomes apparent that there are opportunities for improvement and deeper exploration. One potential extension of this study, which could lead to a more thorough understanding of the topic, involves testing the indexes on specific subsamples, particularly within specialties such as dermatology and reconstructive and cosmetic surgery. These specialties may involve visual representations of procedures, such as “before and after” images [34], which could pose potential threats to e-professionalism.

Improving the quality of external validity assessment can be achieved by incorporating self-evaluation of e-professionalism into the MIMIC model. This addition would enhance the content validity of the model by supplementing existing reflective indicators used in the research. Furthermore, self-evaluation of e-professionalism would serve as a valuable tool for evaluating the nomological network of the instrument. It would provide insights into the direction and strength of correlation among individual indicators of e-professionalism, the e-professionalism indices themselves, and potential predictors for model creation.

Future attempts aimed at measuring e-professionalism could focus on investigating the underlying reasons behind responses such as “I have never been in a situation where this could happen.” It is plausible that a more precise definition of items or the inclusion of specific examples could serve as mechanisms to help respondents differentiate between behaviors they never engage in and those they may never encounter. By refining the clarity and specificity of survey items, researchers can facilitate a more accurate assessment of respondents’ experiences and perceptions related to e-professional behavior. This approach could lead to a deeper understanding of the nuances involved in professional conduct within the context of SNSs.

Conclusions
In this paper, an instrument for measuring the e-professional behavior of MDs and DMDs was developed and validated using the formative approach. Following the validation process, the instrument comprises 19 items, which contribute to the formation of 2 indexes. The first index, focusing on the danger aspect of SNSs, is composed of 14 items that were dichotomized before index construction. The second index, which examines the opportunity aspect of SNSs, is composed of 5 items that were recoded as 4-point items before index construction.

These innovative indexes offer a means to gauge the level of e-professional behavior among MDs and DMDs. This marks the first measure specifically designed to assess the e-professional behavior of MDs and DMDs. The paper demonstrates the feasibility of investigating e-professional behavior using a formative approach, representing an advancement over existing measuring instruments. This approach provides a means to mitigate the impact of cognitive dissonance between attitudes and the actual behavior of MDs and DMDs.

The validation process confirmed that these indexes serve as a robust measure of e-professional behavior. Nevertheless, the instrument has been scrutinized for potential areas of enhancement, and suggestions for improvements have been proposed for future iterations of the instrument.

Acknowledgments
This study was funded by the Croatian Science Foundation under project UIP-05-2017 “Dangers and Benefits of Social Networks: E-Professionalism of Health Care Professionals – SMePROF”. Generative AI was not used in any portion of the manuscript writing.

Data Availability
The data sets used or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File, 23 KB - mededu_v10i1e50156_app1.docx]

Multimedia Appendix 2
Descriptive characteristics of reflective indicators for the MIMIC models of e-professionalism (N=753).
[DOC File, 34 KB - mededu_v10i1e50156_app2.doc]

Multimedia Appendix 3
Type of workplace and specialization status of the respondents.
Multimedia Appendix 4
Intercorrelations of items in the e-professionalism instrument - danger aspect of SNSs (N=753).

Multimedia Appendix 5
MIMIC e-Professionalism models.

Multimedia Appendix 6
Intercorrelations of items in the e-professionalism instrument - opportunity aspect of SNSs (N=753).

References


31. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [DOI: 10.2196/jmir.6.3.e34] [Medline: 15471760]


33. Dangers and Benefits of Social Networks: E-Professionalism of Health Care Professionals – SMePROF. University of Zagreb, School of Medicine. URL: https://mef.unizg.hr/en/znanost-2/istrazi/6.3.1.e34 [Medline: 15471760]


Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABIM</td>
<td>The American Board of Internal Medicine</td>
</tr>
<tr>
<td>CCDM</td>
<td>Croatian Chamber of Dental Medicine</td>
</tr>
<tr>
<td>CHERRIES</td>
<td>Checklist for Reporting Results of Internet E-Surveys</td>
</tr>
<tr>
<td>CHIF</td>
<td>Croatian Health Insurance Fund</td>
</tr>
<tr>
<td>CMC</td>
<td>Croatian Medical Chamber</td>
</tr>
<tr>
<td>DMD</td>
<td>doctor of dental medicine</td>
</tr>
<tr>
<td>HCP</td>
<td>health care professional</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Health Insurance Portability and Accountability Act</td>
</tr>
<tr>
<td>MD</td>
<td>doctor of medicine</td>
</tr>
<tr>
<td>MIMIC</td>
<td>multiple indicators multiple causes</td>
</tr>
<tr>
<td>SNS</td>
<td>social networking site</td>
</tr>
<tr>
<td>VIF</td>
<td>variance inflation factor</td>
</tr>
</tbody>
</table>

Edited by T Leung, T de Azevedo Cardoso; submitted 21.06.23; peer-reviewed by C Asaad, A Jeroncic; comments to author 25.08.23; revised version received 29.09.23; accepted 28.12.23; published 27.02.24.

Please cite as:

Marelić M, Klasnić K, Vukušić Rukavina T
Measuring e-Professional Behavior of Doctors of Medicine and Dental Medicine on Social Networking Sites: Indexes Construction With Formative Indicators
JMIR Med Educ 2024;10:e50156
URL: https://mededu.jmir.org/2024/1/e50156
doi: 10.2196/50156
PMID:38412021

©Marko Marelić, Ksenija Klasnić, Tea Vukušić Rukavina. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Abstract

Medical advice is key to the relationship between doctor and patient. The question I will address is “how may chatbots affect the interaction between patients and doctors in regards to medical advice?” I describe what lies ahead when using chatbots and identify questions galore for the daily work of doctors. I conclude with a gloomy outlook, expectations for the urgently needed ethical discourse, and a hope in relation to humans and machines.

(intext:JMIR Med Educ 2024;10:e50869) doi:10.2196/50869

KEYWORDS
chatbot; ChatGPT; medical advice; ethics; patients; doctors

Introduction

While I strive to provide accurate and helpful information, I am not a substitute for medical advice or professional judgment, and it’s always important for patients and healthcare providers to work together to develop a personalized treatment plan that takes into account a patient’s individual needs and circumstances. [ChatGPT, 2023]

Medical advice (MA) is key to the relationship between doctor and patient. The question I will address is “how may chatbots affect the interaction between patients and doctors in regards to medical advice?” To this end, I shall consider—and go beyond—what was recently outlined regarding MA in “A Conversation With ChatGPT” [1].

Advances in artificial intelligence (AI) and chatbots are changing the world, including medicine [2-4]. ChatGPT is a generative pretrained transformer model based on GPT-3 from OpenAI. Based on word correlations in its 175 billion–parameter database, ChatGPT floods us with meaningful but also nonsensical information.

Concerning the interaction between patients, doctors, and chatbots, I describe what lies ahead when using chatbots and identify many questions for the daily work of doctors. I conclude with a gloomy outlook, expectations for urgently needed ethical discourse [5,6], and a hope in relation to humans and machines [3,7].

Weighing ChatGPT’s Quote

How ChatGPT describes its role [1]—“I am not a substitute for medical advice”—should be a fact. Doctors, as the only authoritative providers of professional MA, must always be in the driver’s seat. Chatbots have the potential to help with the task of contributing general information to an information chain. Importantly, doctors need to review and question all AI output and see if and how it contributes to a patient’s understanding and fits within MA. Depending on the expectations and hopes that ChatGPT raises in patients, this task could become an unprecedented challenge.

With their up-to-date knowledge and medical experience and expertise, doctors need to integrate personal, specific, and general information into their comprehensive MA to the patients. Chatbots are limited to general information stored in databases. Concerningly, ChatGPT invents facts, called a hallucination in
AI [3]. Moreover, ChatGPT can produce nonsensical or “bullshit” [8] information, nicely worded and seemingly justified but disregarding truth and facts—disconcertingly, we do not readily know how often and when ChatGPT offers “bullshit” or nonsense responses.

The Daily Work of Doctors: Question Galore

Nevertheless, ChatGPT will be used by many simply because it is there and seemingly easy and, importantly, free to use. Is it, therefore, likely that we can do without chatbots? No, because society will not abandon ChatGPT or other advanced chatbot tools [3]. The sooner we understand chatbot information for patients, the better. Realistically, ChatGPT is just the tip of an AI iceberg. The “Godfather of AI” [9] Hinton and OpenAI’s chief executive officer Altman [10] have warned forcefully about the speed, impact, and inevitability of AI developments.

Doctors routinely deal with both informed and misinformed patients, fuelled by online health searches (eg, “Dr Google” [11]). Indeed, the internet has become the starting point for many to ask questions about health, disrupting traditional doctor-patient relationships [12] and leading to potential harm from online misinformation [11]. Importantly, neither patients nor doctors should give away too much information when using AI. Even if MA could get better with more details, can we know if this information is being used beyond MA? Indeed, to what extent may creating MA be used as an AI Trojan horse to extract information for other purposes, including business benefits? Which biases go into AI-based medical information, for example, through training data that neither represent the ethnicity nor the financial options of diverse patients? That medically advanced AI may become expensive raises questions of equity: who will have access to these technologies?

What knowledge do doctors need to understand medical AI advice? How can AI-based medical information be used [13], and how do you deal with medical information that AI cannot explain [14]? Could doctors working with chatbot-provided diagnoses and AI-recommended treatments miss the true picture and become overreliant on AI? Who is liable when doctors use AI medical information, and to come full circle, who is liable when they do not [2,15]? Could there come a time when not considering AI such as ChatGPT constitutes less than adequate advice and nonstandard care [15]? Doctors should ask their liability insurer how (ie, under what conditions) and to what extent the insurer covers the use, or nonuse, of AI in practice [15].

Key orientation for interactions between patients, doctors, and chatbots regarding MA can come from physicians’ professional organizations and the US Food and Drug Administration. Similar to practice guidelines [15], recommendations and guardrails for practice-specific medical information via chatbots may have to be developed.

A Gloomy Outlook, Expectations From Much-Needed Ethical Discourse, and a Hope in Relation to Humans and Machines

That ChatGPT “strive(s) to provide accurate and helpful information” [1] has a stale empirical aftertaste. In fact, according to OpenAI, advanced AI [16] will make reviewing chatbot information even more difficult. GPT-4 (eg, in Microsoft Bing and ChatGPT Plus), with 571 times as many learned parameters as GPT-3, has “learned” to deliver incorrect work more convincingly than earlier models. Such mistakes will pose severe problems even if “[ChatGPT] admits these when challenged” [1].

PubMed-listed comparisons between GPT-3 and GPT-4 suggest that the latter may provide more accurate patient information in nuclear medicine [17]. Another study suggested that both free and paid versions of ChatGPT risk providing misleading responses when used without expert MA [18]. Chatbot medical information within at a college reading level suggested that such AI devices may be used supplementarily but not as a primary source for medical information [19], emphasizing the doctor’s key role in MA. More research is needed on MA in numerous medical fields and settings, for numerous applications, and for various populations.

Overall, when AI experts at the University of California, Berkeley explored and discussed the implications of ChatGPT and AI and future challenges in the spring of 2023, there was an explicit call for more ethical considerations [6,20]. Priority safety measures include strict regulations for patient privacy and ethical practices [21]. While the questions above are not exhaustive, it is time to systematically answer them regarding MA and the unavoidable interaction of patients, doctors, and chatbots.

Ultimately, we can only hope that the boundaries between humans and machines [3] will never become so blurred that patients cannot distinguish the MA of a human doctor from the general information provided by ChatGPT [22] or other AI.

Acknowledgments

TCE acknowledges stimulating working conditions as a visiting scholar at the University of California, Berkeley. Support is acknowledged for the article processing charge from the DFG (Deutsche Forschungsgemeinschaft / German Research Foundation, 491454339).

Conflicts of Interest

None declared.
References


Abbreviations

AI: artificial intelligence
MA: medical advice

https://mededu.jmir.org/2024/1/e50869
Generative Language Models and Open Notes: Exploring the Promise and Limitations

Charlotte Blease¹,², PhD; John Torous², MD, MBI; Brian McMillan³, PhD; Maria Hägglund⁴, PhD; Kenneth D Mandl⁵, MD, MPH

¹Department of Women’s and Children’s Health, Uppsala University, Uppsala, Sweden
²Digital Psychiatry, Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, United States
³Centre for Primary Care and Health Services Research, University of Manchester, Manchester, United Kingdom
⁴Medtech Science & Innovation Centre, Uppsala University Hospital, Uppsala, Sweden
⁵Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Charlotte Blease, PhD
Department of Women’s and Children’s Health
Uppsala University
Box 256
Uppsala, 751 05
Sweden
Phone: 46 18 471 00 0
Email: charlotteblease@gmail.com

Abstract

Patients’ online record access (ORA) is growing worldwide. In some countries, including the United States and Sweden, access is advanced with patients obtaining rapid access to their full records on the web including laboratory and test results, lists of prescribed medications, vaccinations, and even the very narrative reports written by clinicians (the latter, commonly referred to as “open notes”). In the United States, patient’s ORA is also available in a downloadable form for use with other apps. While survey studies have shown that some patients report many benefits from ORA, there remain challenges with implementation around writing clinical documentation that patients may now read. With ORA, the functionality of the record is evolving; it is no longer only an aide memoire for doctors but also a communication tool for patients. Studies suggest that clinicians are changing how they write documentation, inviting worries about accuracy and completeness. Other concerns include work burdens; while few objective studies have examined the impact of ORA on workload, some research suggests that clinicians are spending more time writing notes and answering queries related to patients’ records. Aimed at addressing some of these concerns, clinician and patient education strategies have been proposed. In this viewpoint paper, we explore these approaches and suggest another longer-term strategy: the use of generative artificial intelligence (AI) to support clinicians in documenting narrative summaries that patients will find easier to understand. Applied to narrative clinical documentation, we suggest that such approaches may significantly help preserve the accuracy of notes, strengthen writing clarity and signals of empathy and patient-centered care, and serve as a buffer against documentation work burdens. However, we also consider the current risks associated with existing generative AI. We emphasize that for this innovation to play a key role in ORA, the cocreation of clinical notes will be imperative. We also caution that clinicians will need to be supported in how to work alongside generative AI to optimize its considerable potential.

(JMIR Med Educ 2024;10:e51183) doi:10.2196/51183

KEYWORDS
ChatGPT; generative language models; large language models; medical education; Open Notes; online record access; patient-centered care; empathy; language model; online record access; documentation; communication tool; clinical documentation

Introduction

Patient online record access (ORA) is growing globally [1]. Access includes test and laboratory results, secondary or hospital care letters, lists of prescribed medications, and the narrative reports written by clinicians after visits (the latter referred to as “open notes”). Already, patients across an estimated 30 countries can access some of their records via secure web portals including
health apps. In some countries, this innovation is advanced [1]. Since 2021, the federally enacted 21st Century Cures Act in the United States mandated that providers offer all patients access to download their electronic health records without charge [2]. In the Nordic countries, ORA has been implemented incrementally, starting around 2010 [3]. The Finnish patient portal OmaKanta was rolled out with stepwise implementation of functionality between 2010 and 2015 [4]. Patients in Sweden first obtained ORA in one of 21 regions in 2012 [5] with nationwide implementation achieved by 2018. Implementation in Norway began in 2015, reaching patients in 3 out of 4 regions by 2019 [6]. In England, from October 31, 2023, it is mandatory for general practitioners to offer ORA to their adult patients, albeit on a prospective basis [7].

Patients with access to their records report using them to become more involved in their care, to follow up on doctors’ visits, and to obtain an overview of their test results and treatment history [3,8,9]. Multiple surveys show that patients using ORA are positive about the experience after reading their notes. They report many benefits including understanding their care plans better [9], improved communication with and greater trust in their provider [10], and feeling more in control of their health and care [6,8], including doing a better job taking their medications [11,12].

Despite the patient benefits with ORA, challenges with their implementation in clinical practice remain. In this viewpoint paper, we identify outstanding concerns with ORA, which encompass a range of unintended consequences for clinician work burdens, and for the substantial task of conveying bespoke, compassionate, and understandable information to each unique patient who accesses their records. Currently, it has been proposed that a range of targeted patient training and medical education strategies may be sufficient to resolve at least some of these challenges [13-17]. We believe that such interventions are valuable; however, in this viewpoint paper, we explain why the ambitions of such training interventions may be limited.

As a solution, we explain why the use of generative artificial intelligence (AI) may offer more tangible long-term promise than clinician training alone in helping to resolve problems with ORA implementation. While generative AI itself is not new, recent technical advances and the increased accessibility of large language models (LLMs; GPT-4 by OpenAI, LLaMA by Meta, and PaLM2 by Google) have made clinical use increasingly feasible. LLMs are an application of generative AI technology, often defined as machine learning algorithms that can recognize, summarize, and generate content based on training on large data sets. Unlike search engines, which offer pages of internet links in response to typed queries, generative LLMs such as GPT-4 simulate well-reasoned answers couched as conversations. In addition, these models can “remember” previous prompts, helping to build up the perception of dialogic exchange. We review the strengths and limitations of generative AI and emphasize that for this innovation to play a key role in ORA, it will be imperative for humans to be involved as overseers of computer input.

Current Challenges With Open Notes

Evolving Functionality of Records

Guidelines, such as those issued by the British General Medical Council, state that clinicians should keep clear, accurate, contemporaneous records that include “…any minor concerns, and the details of any action you have taken, information you have shared and decisions you have made relating to those concerns” [18]. In the era of ORA, clinicians will also need to consider if what they write will be understandable, accessible, and supportive for patients [19]. With the knowledge that patients will read what they write, the functionality of the record is evolving, and this incurs changes with respect to how clinical information is documented [20,21]. Clinicians must uphold the original functionality of the record—documenting the patient’s medical information in clinical detail, but also communicating this information to the patient. With respect to the latter function, it is argued that for records to be understandable and acceptable to a lay audience, clinicians should ideally remove or explain medical acronyms, omit medical vernacular that may be perceived as offensive (such as “patient denies” or “patient complains of”), and strive to convey information in a manner that it is straightforward, comprehensive, and empathic in tone [14]. This is not an easy undertaking for clinicians tasked with pitching information at a literacy level that accommodates diverse patient populations while maintaining the clinical utility of records and adequately serving their medicolegal functions. Indeed, whether such dual functionality is even possible has been questioned [22].

Documentation Changes

To date, it is unclear whether ORA diminishes the clinical value of documentation [19,23]. However, there is evidence that clinicians may be undermining the accuracy or completeness (or both) of their records, perhaps in attempts to reduce patient anxieties, minimize follow-up contact, or reduce the likelihood of potential complaints [24,25]. For example, in the largest study conducted on clinicians’ experiences of open notes, a 3-center study at 3 diverse health systems in the United States (1628 of 6054, 27% clinicians responded), DesRoches et al [26] found that around 1 in 4 physicians admitted that they changed how they wrote differential diagnoses (23%, n=176), though the nature of these changes is not understood. More worryingly, more than 1 in 5 physicians (22%, n=168) believed that their notes were now less valuable for other clinicians [26].

Conceivably, other changes following implementation of ORA might be more positive. In the study by DesRoches et al [26], 22% (n=166) of physicians reported changes to the use of a partnering language, and 18% (n=139) of them reported changes to how they used medical jargon or acronyms. However, it remains unknown whether such changes improve the comprehensibility of clinical records among patients or whether amendments come with a trade-off in terms of documentation quality.

With ORA, there is also the potential for notes to convey bias of stigmatizing language. For example, in the United States, recent linguistic analysis studies have shown that negative patient descriptors in notes are considerably more common for
non-Hispanic black patients and for patients with diabetes, those with substance use disorders, and those with chronic pain [27,28]. It is unclear whether with the knowledge patients may now read what they write, the use of stigmatizing language among these patient populations is being effectively omitted and “cleaned up” by clinicians.

**Work Burdens**

Time spent on documentation and patient portal messages remains a growing cause of clinician dissatisfaction and burnout [29]. The impact is exacerbated for clinicians with lower levels of digital competencies, and this “technostress” has been found to directly correlate with burnout [30]. Even tech-savvy young resident physicians have reported the use of the electronic health record as a leading cause of burnout [31]. In the United States, the study by DesRoches et al [26] on clinicians’ experiences, 37% (n=292) of physicians reported spending more time writing notes after patient access was enabled.

Few studies have explored objective measures of the impact of ORA, however, where these measures have been implemented, some of them signal potential for increased patient contact. For example, Mold et al [32] found that the provision of ORA in primary care settings resulted in a moderate increase in email traffic from patients, with no change in telephone contact and variable changes to face-to-face contact. A recent Canadian study found that registration with a primary care web-based portal was associated with an increase in the number of visits to physicians, calls to practice triage nurses, and an increase in clerical workload [33]. Another recent study at an academic medical center in the United States reported a doubling in the number of messages sent by patients within 6 hours after ORA was implemented [34]. It seems reasonable to postulate that at least some of this increased contact may be driven by patients who desire clarifications about diagnoses, results, or other information that is documented in their records.

**Currently Proposed Solutions**

To encourage confidence with ORA and to overcome some of these challenges, targeted educational programs have been proposed. Among them are short lists of tips and advice to clinicians, and brief web-based training interventions [13,14,24]. More recently, some medical schools have taken this further. For example, Harvard Medical School has embedded within its curriculum practical training in how to write notes that patients will read [16], and similar work is underway in England [35]. The expressed aim of such training programs is to support clinicians writing notes that are more suited to a readership like health literacy needs. We suspect that the trade-off may lead to the cause are sufficient to counter the hidden curriculum or to neutralize the formation of documentation habits that may not be in keeping with the ORA mandate whereupon clinical notes may now be read by patients and caregivers.

Other recommendations that clinicians should remove all acronyms and medical jargon may present practical dilemmas for upholding the quality of documentation. Aside from extra time spent typing documentation, the capacity to shift from expert to patient perspectives poses unappreciated difficulties. Undoubtedly, many clinicians, as domain experts, might not always fully appreciate when they are using specialist or technical language, nor do they have the attendant skills to convey what they know to patients in an understandable way—a cluster of problems collectively referred to as the “curse of expertise” [38]. Using imprecise language may also have future medical consequences and might result in harm if later clinicians misinterpret what was written [39].

Relatedly, it seems a significant request that clinicians write notes that are bespoke for every patient’s level of health literacy. Yet, each person who attends a clinical visit will have specific needs that may now be read by patients and caregivers. While often considered a “soft skill,” the adoption of empathetic, encouraging, and supportive language might be a taller order than is frequently assumed. For example, psychologists report that negative biases can curb expressions of empathy [40-44]. Studies show that empathy can be influenced by patients’ race or ethnicity and may be diminished among people presenting with disabilities or already stigmatized conditions [40-44]. Making matters worse, self-inspection may be a particularly weak tool for clinicians to excavate and monitor their own prejudices [45]. Furthermore, the demand that clinicians tailor their notes in ways that are optimized to every patient’s understanding and their emotional needs may lead to not only increased workload but also higher risk of burnout [46].
So far, no objective measures have assessed whether targeted training strategies are effective at improving clinical documentation in terms of preserving medical detail and utility, strengthening patient understanding and patients’ perceptions of clinician support and empathy. We emphasize that while commonly used in training evaluation, self-report surveys will not be sufficient to establish whether educational interventions work in terms of both preserving the detail in clinical notes and supporting patient understanding.

Finally, perhaps most crucial of all, and as already noted, it is unclear whether narrative notes can ever uphold a genuine dual functionality targeting the needs of both clinician and patient readerships [22]. Conceivably, both needs are incommensurable and there will always be a trade-off in detail and understanding should the patient, or the clinician, be given primacy as target reader.

Generative Language Models Writing Clinical Notes

Strengths of Generative AI

Doctors strongly desire support with documentation including note writing with surveys showing that they forecast a role for AI in assisting in these tasks [47,48]. Because of their promise with respect to administrative and documentation tasks in health care contexts, LLMs have been described as “the ultimate paperwork shredder” [49]. Owing to the sheer speed and scope of information upon which they draw, LLMs hold considerable potential in generating up-to-date, comprehensive clinical information for patients [50]. This makes the approach particularly promising in generating detailed narrative explanations and summaries of visit encounters. This may help to reduce work burdens on physicians tasked with writing clinical notes.

Another striking strength of LLMs is their capacity to write responses in a requested style or by adopting a specific tone or conversational emphasis. This makes LLMs particularly promising in assisting with writing notes that omit the use of medical jargon or acronyms that are suitable for patients with different levels of health literacy, or among speakers of languages that differ from their provider’s language. This capacity may also help avoid the extra burdens on clinicians attempting to document notes that are tailored to the highly diverse range of unique patient readers.

Preliminary research also suggests that LLMs may help with writing consistently sensitive or empathic notes. In 2023, a highly publicized study suggested that ChatGPT may have better bedside manners than actual human doctors [51]. A team compared written responses of doctors and ChatGPT offered to patients’ real-world health queries using Reddit’s AskDocs forum, where nearly half a million people post their medical problems and verified and credentialled clinicians offer suggestions. On average, ChatGPT responses were 4 times longer than doctors’ replies. A panel of health care professionals—blinded to who or what did the writing—preferred ChatGPT’s responses nearly 80% of the time. The panel ranked chatbot answers as being of significantly higher quality than web-based posts reportedly from doctors; they also judged these reported web-based doctors’ answers as more unacceptable responses to patients. ChatGPT’s responses were rated as “good” or “very good” nearly 4 times more often than those written by the reported web-based doctors, and ChatGPT’s responses were rated as almost 10 times more empathic than those by the reported web-based doctors. At the other end of the scale, these web-based physicians’ replies were perceived to lack empathy approximately 5 times more often than responses produced by ChatGPT.

Limitations of Generative AI

Despite their potential, LLMs have multiple limitations. The nature of the data sets the models are trained on is critical, as it will determine the scope and nature of responses possible. Of special relevance here, none of the easily accessible LLMs have yet been trained on medical texts and thus lack the core substrate to generate the most appropriate responses. Any bias in the source the models are trained on will also be reflected in answers or text provided. Thus, while a study in March 2023 showed that ChatGPT (version 3) Could pass the United States Medical Licensing Examination [52], the authors of the study noted that to truly assess the potential of such LLMs, there is a need for “controlled and real-world learning scenarios with students across the engagement and knowledge spectrum.” Still, the results of that study were acknowledged by the American Medical Association, which noted that it intends to begin considering how tools such as ChatGPT need to be incorporated into the education process [53].

Indeed, the full extent to which LLMs embed discriminatory biases has not been fully explored. However, it would be surprising if these models did not replicate many of the same biases that already exist in clinical research, and consequently medical education, in part because of the underrecruitment of women, racial and ethnic minorities, and older people. Such skewing is already recognized as a source of disparity with the potential to perpetuate errors or misjudgments in clinical decisions [54-58]. Studies suggest that gender and racial biases are indeed coded into LLMs [59]. It remains unknown whether the potential for such discriminatory errors might prove worse than today with standard human-mediated care; however, some preliminary research suggests that negative stereotyping may be compounded by LLMs [60].

Another concern is the lack of consistency in responses proffered by LLMs. Inputting the same question to GPT-4, for example, rarely elicits the same response. Of course, human responses are rarely consistent as well; however, the extent to which generative AI, relying on LLMs, offers the same level of reliable outputs is uncertain. This is a particular concern given that LLMs are prone to yield falsehoods—a phenomenon referred to as “hallucination.” Moreover, the persuasive conversational tone of LLMs such as GPT-4 means that narrative responses may appear compelling but factually incorrect.

The extent to which doctors may already be adopting generative AI tools, such as OpenAI’s ChatGPT, is not yet known. In the United States, under the 1996 Health Insurance Portability and Accountability Act (HIPAA), which established national standards in the United States to protect patients’ health...
information from being shared by “covered entities”—that is, providers—to other third parties. Therefore, the use of OpenAI, for example, is precluded under the HIPAA. At the time of writing, in the most common use cases, uploading patient details to versions of generative AI would breach patient trust and medical confidentiality due to privacy concerns.

However, the scope for this is quickly changing. Epic—the US software giant which has an estimated 78% of the share of hospital medical record use in the United States [61]—is currently piloting the integration of HIPAA-compliant GPT services [62]. In addition, an Azure HIPAA–compliant GPT-4 service already exists [63]. Voice-to-text clinical note generation products now represent a growing space in health care. For example, a new app called Ambient Experience from Nuance can listen to the physician’s conversation and, using ChatGPT (version 4), help create the clinical note that is ready for physicians to review [64]. In the United States, such capacities are set to become embedded into electronic health systems, signaling revolutionary changes in medical documentation practices.

Clinicians and Computers as Coauthors

Combined, the aforementioned discourse suggests that LLMs are far from ready to disintermediate clinicians when it comes to writing clinical notes. We argue that the innovation will play a key role if humans are involved. Thus, this promise could be harnessed if clinicians oversee the cocreation of clinical documentation. In this scenario, LLMs might offer initial draft documentation, which, crucially, should be supervised, and edited by clinicians whose key role in documentation will be to keep a check and balance on the current limitations with these models.

Considering the scope of generative AI, we therefore propose that current training interventions might be constructively adapted to better prepare clinicians to oversee the writing of patient-facing clinical documentation, for example, by editing and checking the quality of clinical information constructed by generative AI and reviewing the sensitivity of the language used. Preliminary studies already show that when humans collaborate with LLMs to coproduce replies to patients, this can enhance patients’ ratings of levels of empathy compared with human-only produced responses [65]. Such partnership could offer a more robust and safe form of documentation quality control—one that could potentially avoid the work burdens associated with documentation burdens and, therefore, the potential for burnout from ORA. We emphasize, however, that training should reinforce the importance of using generative AI as an assistant narrative scribe and not as a substitute for writing notes.

Furthermore, if health systems adopt this approach, we suggest that 2 (or even multiple) versions of clinical documentation may be feasible. Using LLMs, there is scope to not only a complete medical narrative pitched at the level of the domain expert or specialist, but also to document notes couched at the level of health literacy, language, and empathy of the individual patient who might be reading them. This could help overcome the current dilemma of documenting information in a way that is accessible for patients, but which does not diminish the clinical detail for health professionals.

Future Research Directions

Many research questions could usefully explore generative AI in cowriting clinical notes, especially dual-purpose documentation for both patients and clinicians. We suggest a few novel directions. First, qualitative studies could usefully explore how successfully generative AI translates clinical documentation into patient-friendly language. For example, studies could examine the accuracy and fidelity of generative AI in translating acronyms or other medical jargon, as well as the understandability of the notes, and the level of empathy embedded in patient-facing documentation. Second, experimental studies could probe whether documentation embeds biases or a higher likelihood of containing stigmatizing language for different patient demographics or health conditions. Third, pilot studies could help determine the satisfaction and administrative work burden of dual documentation among clinicians.

Conclusions

Generative AI is ready for mass use when it comes to writing or cowriting clinical notes, and its potential is enormous. We emphasize, however, that there remain evidence-based risks associated with existing generative AI, which relate to inconsistencies, errors, and hallucinations and the real potential to embed harmful biases in documentation. If carefully implemented, in the long term, doctors who write documentation using generative AI may do a better job of adapting to the evolving functionality of the electronic records than doctors who do not. This adoption may address the potential risk of “dumbing down” clinical documentation while conveying understandable and empathetic information to patients using plain and sensitive language. We also forecast that doctors who cowrite their documentation with LLMs will experience fewer work burdens.

Conflicts of Interest

JT is the Editor-in-Chief of JMIR Mental Health. The other authors declare no conflicts of interest.

References


64. DePeau-Wilson M. Clinical Note Writing App Powered by GPT-4 Set to Debut This Year. MedPage Today. 2023. URL: http://tinyurl.com/2sv8dfj4 [accessed 2023-12-15]


Abbreviations

AI: artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
ORA: online record access

Edited by K Venkatesh; submitted 31.07.23; peer-reviewed by B Senst, J Schwarz, R Marshall; comments to author 28.09.23; revised version received 30.09.23; accepted 10.11.23; published 04.01.24.

Please cite as:
Blease C, Torous J, McMillan B, Hägglund M, Mandl KD
Generative Language Models and Open Notes: Exploring the Promise and Limitations
JMIR Med Educ 2024;10:e51183
URL: https://mededu.jmir.org/2024/1/e51183
doi: 10.2196/51183
PMID: 38175688
Original Paper

Pure Wisdom or Potemkin Villages? A Comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 Style Questions: Quantitative Analysis

Leonard Knoedler¹, MD; Michael Alfertshofer², MD; Samuel Knoedler¹,³, BSc; Cosima C Hoch⁴, BSc; Paul F Funk⁵, BSc; Sebastian Cotofana⁶,⁷, MD, PhD; Bhagvat Maheta⁸, BS; Konstantin Frank⁹, MD; Vanessa Brébant¹, MD; Lukas Prantl¹, MD, PhD; Philipp Lamby¹, MD, PhD

¹Department of Plastic, Hand and Reconstructive Surgery, University Hospital Regensburg, Regensburg, Germany
²Division of Hand, Plastic and Aesthetic Surgery, Ludwig-Maximilians University Munich, Munich, Germany
³Division of Plastic Surgery, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, United States
⁴Department of Otolaryngology, Head and Neck Surgery, School of Medicine, Technical University of Munich, Munich, Germany
⁵Department of Otolaryngology, Head and Neck Surgery, University Hospital Jena, Friedrich Schiller University Jena, Jena, Germany
⁶Department of Dermatology, Erasmus Hospital, Rotterdam, Netherlands
⁷Centre for Cutaneous Research, Blizard Institute, Queen Mary University of London, London, United Kingdom
⁸College of Medicine, California Northstate University, Elk Grove, CA, United States
⁹Ocean Clinic, Marbella, Spain

Corresponding Author:
Leonard Knoedler, MD
Department of Plastic, Hand and Reconstructive Surgery
University Hospital Regensburg
Franz-Josef-Strauß-Allee 11
Regensburg, 93053
Germany
Phone: 49 151 44824958
Fax: 49 941 944 6899
Email: leonardknoedler@t-online.de

Abstract

Background: The United States Medical Licensing Examination (USMLE) has been critical in medical education since 1992, testing various aspects of a medical student’s knowledge and skills through different steps, based on their training level. Artificial intelligence (AI) tools, including chatbots like ChatGPT, are emerging technologies with potential applications in medicine. However, comprehensive studies analyzing ChatGPT’s performance on USMLE Step 3 in large-scale scenarios and comparing different versions of ChatGPT are limited.

Objective: This paper aimed to analyze ChatGPT’s performance on USMLE Step 3 practice test questions to better elucidate the strengths and weaknesses of AI use in medical education and deduce evidence-based strategies to counteract AI cheating.

Methods: A total of 2069 USMLE Step 3 practice questions were extracted from the AMBOSS study platform. After including 229 image-based questions, a total of 1840 text-based questions were further categorized and entered into ChatGPT 3.5, while a subset of 229 questions were entered into ChatGPT 4. Responses were recorded, and the accuracy of ChatGPT answers as well as its performance in different test question categories and for different difficulty levels were compared between both versions.

Results: Overall, ChatGPT 4 demonstrated a statistically significant superior performance compared to ChatGPT 3.5, achieving an accuracy of 84.7% (194/229) and 56.9% (1047/1840), respectively. A noteworthy correlation was observed between the length of test questions and the performance of ChatGPT 3.5 ($ρ$=−0.069; $P$=.003), which was absent in ChatGPT 4 ($P$=.87). Additionally, the difficulty of test questions, as categorized by AMBOSS hammer ratings, showed a statistically significant correlation with performance for both ChatGPT versions, with $ρ$=−0.289 for ChatGPT 3.5 and $ρ$=−0.344 for ChatGPT 4. ChatGPT 4 surpassed ChatGPT 3.5 in all levels of test question difficulty, except for the 2 highest difficulty tiers (4 and 5 hammers), where statistical significance was not reached.

https://mededu.jmir.org/2024/1/e51148

JMIR Med Educ 2024 | vol. 10 | e51148 | p.164

(page number not for citation purposes)
Conclusions: In this study, ChatGPT 4 demonstrated remarkable proficiency in taking the USMLE Step 3, with an accuracy rate of 84.7% (194/229), outshining ChatGPT 3.5 with an accuracy rate of 56.9% (1047/1840). Although ChatGPT 4 performed exceptionally, it encountered difficulties in questions requiring the application of theoretical concepts, particularly in cardiology and neurology. These insights are pivotal for the development of examination strategies that are resilient to AI and underline the promising role of AI in the realm of medical education and diagnostics.

Methods

Access to Question Bank and Data Entry Procedure
From June 12, 2023, to June 19, 2023, we obtained access to the AMBOSS question bank [13]. Within this time frame, we collected a total of 1840 practice questions specifically designed for the USMLE Step 3 exam. Before initiating our study, we acquired official permission from AMBOSS (AMBOSS GmbH) to use their USMLE Step 3 question bank for research purposes. To ensure the reliability of our data, 2 examiners (MA and LK) cross-checked the question inputs randomly to confirm that none of the answers were indexed on Google before June 19, 2023. Many USMLE questions are on the internet, including USMLE sample questions as well as a few AMBOSS questions; however, we ensured that those questions were not included in this analysis to minimize the risk of prior memorization of the questions by ChatGPT. July 19, 2023, was chosen since it represents the most recent accessible date within the training data set of ChatGPT. There are many forms of AI versions with capabilities to answer USMLE Step 3 practice test questions; however, ChatGPT is the most widely used AI at the time of this study, making it the best fit for our study.

Question Screening and Categorization
To maintain the quality of our sample questions, we subjected all test questions to independent screening by 4 examiners (MA, SK, CCH, and LK). Questions containing clinical images and photographs were excluded from the study, resulting in the removal of 229 image-based questions. Subsequently, the remaining 1840 test questions were classified based on their respective specialties, using the categorization provided by AMBOSS. All questions included in our study followed a multiple-choice single-answer format. The questions used for both ChatGPT 3.5 and ChatGPT 4 were matched for content and difficulty based on the standardized definitions provided by the AMBOSS question bank to ensure consistent analysis between both AI versions.

Comparison of ChatGPT Versions and Analysis of Question Stems
To evaluate any performance differences between ChatGPT 3.5 and ChatGPT 4, we conducted a subgroup analysis specifically focusing on ChatGPT 4. Additionally, we analyzed the question stems of both ChatGPT 3.5 and ChatGPT 4, specifically looking for specific buzzwords related to diagnostic methods and patient information, such as “Ultrasound,” “SEROLOGY,” and “Nicotine Abuse.” These particular words and phrases may suggest one...
answer over another and thus are essential for test-taking. For example, if the question states “Nicotine Abuse,” which is suggestive of cigarette or tobacco use, the patient in the question stem is more likely to have cancer. The purpose of this analysis was to identify any variations in accuracy based on the presence of these factors. Furthermore, we assessed performance differences between ChatGPT 3.5 and ChatGPT 4 based on the length of the test questions.

Assessment of Question Difficulty

To assess the difficulty of the test questions, we used the proprietary rating system of the AMBOSS question bank. This system assigns a difficulty level to each question based on a scale of 1 to 5 hammers. A rating of 1 hammer corresponds to the easiest 20% of questions, while 5 hammers indicate the most challenging 5% of questions.

Data Entry Process

One examiner (MA) manually inputted the test questions into ChatGPT. The questions were transcribed verbatim from the AMBOSS question bank, preserving the original text and answer choices. To ensure the integrity of ChatGPT’s performance, no additional prompts were introduced intentionally by the authors, thereby minimizing the potential for systematic errors. Each question was treated as a separate chat session in ChatGPT to minimize the impact of memory retention bias. As an example, the following provides a standard test question from the category “Competency: Patient Care Content Area: General Principles”:

What is the most suitable course of action to take next in the case of a 54-year-old man, previously in good health, who presents to the emergency department after being bitten by a stray dog in South America? The bite punctured his right leg, but he has diligently cleaned the wound daily with soap and peroxide. The patient is not experiencing pain, fever, or chills, and his vital signs are normal. The examination reveals healing puncture wounds with minimal redness, and there is no fluctuation or palpable lymph nodes in the groin. The patient had a tetanus booster vaccination three years ago.

(A) Provide rabies vaccination
(B) Administer tetanus immune globulin
(C) Request cerebrospinal fluid analysis
(D) Order an MRI [magnetic resonance imaging] scan of the brain and spinal cord
(E) No immediate action is required at this time

Recording and Evaluation of ChatGPT Responses

The answers generated by ChatGPT were documented and incorporated into the corresponding AMBOSS USMLE Step 3 practice question. Subsequently, we systematically gathered and recorded information regarding the accuracy of these responses in a separate data spreadsheet.

Statistical Analysis

We used the Pearson chi-square test to determine differences in question style and categories. Bivariate correlation analysis between ChatGPT performance, test question length, and difficulty was conducted using the Spearman correlation coefficient ($\rho$). IBM SPSS Statistics 25 (IBM Corp) was used for statistical analysis, and a 2-tailed $P$ value $\leq .05$ was considered statistically significant.

Results

General Test Question Characteristics and Performance Statistics

The overall accuracy of ChatGPT 3.5 for USMLE Step 3 was 56.9% (1047/1840), while ChatGPT 4 answered 84.7% (194/229) of test questions correctly ($P$ <.001). Specialty-specific number of test questions and performance scores are presented in Tables 1 and 2. ChatGPT 3.5 received the greatest number of questions on the nervous, cardiovascular, and gastrointestinal systems, while ChatGPT 4 received the greatest number of questions on behavior health, the female reproductive system, as well the blood and lymphatic system. When considering the accuracy of ChatGPT based on the category of questions, ChatGPT 3.5 performed the best on behavioral health, multisystem processes and disorders, and pregnancy-related questions. On the other hand, ChatGPT 4 had the greatest accuracy on questions related to the endocrine and musculoskeletal systems as well as biostatistics and multisystem processes and disorders.
Table 1. The number of test questions answered by ChatGPT 3.5 and its performance, stratified by questions category (N=1840).

<table>
<thead>
<tr>
<th>Question category</th>
<th>Test questions answered, n</th>
<th>Correct questions, n/N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male reproductive system</td>
<td>28</td>
<td>17/28 (60.1)</td>
</tr>
<tr>
<td>General principles and foundational science</td>
<td>29</td>
<td>16/29 (55.2)</td>
</tr>
<tr>
<td>Immune system</td>
<td>40</td>
<td>25/40 (62.5)</td>
</tr>
<tr>
<td>Skin and subcutaneous tissue</td>
<td>72</td>
<td>39/72 (54.2)</td>
</tr>
<tr>
<td>Renal and urinary systems</td>
<td>72</td>
<td>39/72 (54.2)</td>
</tr>
<tr>
<td>Biostats and epidemiology</td>
<td>87</td>
<td>45/87 (51.7)</td>
</tr>
<tr>
<td>Female reproductive system and breast</td>
<td>88</td>
<td>48/88 (54.5)</td>
</tr>
<tr>
<td>Musculoskeletal system</td>
<td>94</td>
<td>56/94 (58.5)</td>
</tr>
<tr>
<td>Endocrine system</td>
<td>103</td>
<td>58/103 (56.3)</td>
</tr>
<tr>
<td>Blood and lymphoreticular system</td>
<td>105</td>
<td>55/105 (52.4)</td>
</tr>
<tr>
<td>Pregnancy, childbirth, and puerperium</td>
<td>111</td>
<td>66/111 (59.5)</td>
</tr>
<tr>
<td>Behavioral health</td>
<td>115</td>
<td>73/115 (63.5)</td>
</tr>
<tr>
<td>Multisystem processes and disorders</td>
<td>122</td>
<td>73/122 (59.8)</td>
</tr>
<tr>
<td>Respiratory system</td>
<td>130</td>
<td>71/130 (54.6)</td>
</tr>
<tr>
<td>Social sciences</td>
<td>141</td>
<td>86/141 (61.0)</td>
</tr>
<tr>
<td>Gastrointestinal system</td>
<td>156</td>
<td>87/156 (55.8)</td>
</tr>
<tr>
<td>Cardiovascular system</td>
<td>161</td>
<td>89/161 (55.3)</td>
</tr>
<tr>
<td>Nervous system and special senses</td>
<td>186</td>
<td>104/186 (55.9)</td>
</tr>
</tbody>
</table>

Table 2. The number of test questions answered by ChatGPT 4 and its performance, stratified by questions category (N=229).

<table>
<thead>
<tr>
<th>Question category</th>
<th>Test questions answered, n</th>
<th>Correct questions, n/N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endocrine system</td>
<td>1</td>
<td>1/1 (100)</td>
</tr>
<tr>
<td>Biostats and epidemiology</td>
<td>14</td>
<td>13/14 (92.3)</td>
</tr>
<tr>
<td>General principles and foundational science</td>
<td>17</td>
<td>14/17 (82.4)</td>
</tr>
<tr>
<td>Multisystem processes and disorders</td>
<td>17</td>
<td>15/17 (88.2)</td>
</tr>
<tr>
<td>Pregnancy, childbirth, and puerperium</td>
<td>19</td>
<td>15/19 (79.0)</td>
</tr>
<tr>
<td>Gastrointestinal system</td>
<td>21</td>
<td>18/21 (85.7)</td>
</tr>
<tr>
<td>Cardiovascular system</td>
<td>21</td>
<td>15/21 (71.4)</td>
</tr>
<tr>
<td>Nervous system and special senses</td>
<td>21</td>
<td>15/21 (71.4)</td>
</tr>
<tr>
<td>Blood and lymphoreticular system</td>
<td>23</td>
<td>20/23 (87.0)</td>
</tr>
<tr>
<td>Female reproductive system and breast</td>
<td>23</td>
<td>20/23 (87.0)</td>
</tr>
<tr>
<td>Behavioral health</td>
<td>24</td>
<td>21/24 (87.5)</td>
</tr>
</tbody>
</table>

Test Question Length and ChatGPT Performance Scores
The mean character count was 1078 (SD 308). Test question length was significantly correlated with the performance of ChatGPT 3.5 (p=.069; P=.003) while not yielding significance for ChatGPT 4 (P=.87). For ChatGPT 3.5, the mean number of characters was 1062 (SD 310) for correct answers versus 1100 (SD 304) for falsely answered questions (P=.009). However, the mean character count was comparable for test questions answered by ChatGPT 4 (mean correct answers 1068, SD 274 vs mean false answers 1056, SD 233; P=.80).

Test Question Difficulty and the Performance of ChatGPT
Question distribution and performance scores sorted by level of test question difficulty are illustrated in Figure 1. Test question difficulty, defined by AMBOSS hammer categorization, and the performance of ChatGPT 3.5 were significantly correlated (p=.0289; P<.001). This was reproducible in ChatGPT 4 (p=.344; P<.001). ChatGPT 4 statistically significantly outperformed ChatGPT 3.5 for each hammer category except for the 4- and 5-hammer test difficulty levels. For 1-, 2-, and 3-hammer questions, ChatGPT 4 had a statistically significant increase in accuracy compared to
ChatGPT 3.5 ($P = .04$; $P = .02$; and $P = .03$; respectively). For the most difficult questions, ChatGPT 4 still had greater accuracy than ChatGPT 3.5; however, there was no statistical significance shown. The percentage of correct responses from ChatGPT 3.5 versus ChatGPT 4 sorted by specialty is illustrated in Figure 2.

Relative to ChatGPT 3.5, ChatGPT 4 performed better on questions from every specialty category. The biggest differences in accuracy were in biostatistics, epidemiology, the endocrine system, and the musculoskeletal system.

Figure 1. Question distribution and performance scores sorted by level of test question difficulty.
Figure 2. Percentage of correct responses from ChatGPT 3.5 versus ChatGPT 4.0, sorted by specialty.

Buzzwords and the Performance of ChatGPT
ChatGPT 4 performed significantly better on ultrasound-related questions ($P=.04$), while ChatGPT 3.5 answered significantly more questions correctly if they contained serology- or smoking-related information ($P=.008$ and $P=.03$, respectively). Performance scores of ChatGPT 3.5 versus ChatGPT 4 sorted by buzzwords are depicted in Figure 3. Overall, ChatGPT 4 outperformed ChatGPT 3.5, regardless of whether the question included buzzwords.
**Discussion**

**Principal Findings**

This investigation was designed to empirically evaluate and contrast the competencies of the 2 most contemporary iterations of the AI-powered large language model, ChatGPT, in relation to their performance in taking the USMLE Step 3. An aggregate of 1840 representative practice questions, derived from the AMBOSS question bank, were presented to ChatGPT version 3.5. The model delivered an overall accuracy rate of 56.9% (1047/1840). In juxtaposition, ChatGPT version 4 was assessed using a subset of 229 practice questions and achieved an overall accuracy rate of 84.7% (194/229). This difference in performance is both statistically and practically significant. Achieving a score of 84.7%, ChatGPT 4 falls within the top 10% of all test takers. In contrast, a score of 56.9% places ChatGPT 3.5 near the passing threshold. This significant difference provides empirical evidence of the substantial enhancements and refinements embedded within ChatGPT 4 and elucidates the leap in proficiency this iteration has attained, pushing the boundaries of AI capabilities in medical knowledge comprehension and application.

While ChatGPT 3.5 hovered around the approximate passing threshold of 60%, ChatGPT 4 not only passed the examination but merely excelled at it. According to the score interpretation guide provided by the National Board of Medical Examiners, an accuracy rate of 84.7% approximates placement within the 90th to 92nd percentile [14]. This signifies that ChatGPT 4 would be situated among the elite stratum, encompassing the top 10% of USMLE Step 3 candidates. The impressive escalation in performance exhibited by ChatGPT 4 makes the delineation of strengths and limitations difficult [15]. The model’s evolution seems to have attenuated discernible weaknesses, indicating a more well-rounded overall proficiency in the medical domain [12].

However, nothing is perfect. Although ChatGPT 4 accesses detailed, comprehensive, and up-to-date knowledge bases to optimize its response patterns, we could reveal minor...
performance weak points. We found that ChatGPT 4 was more prone to errors when answering test questions on cardiology (mean test accuracy: n=89, 71.4% vs n=15, 84.7% correct questions) and neurology (mean test accuracy: n=104, 71.4% vs n=15, 84.7% correct questions). Interestingly, these subjects often test the examinee’s transfer knowledge skills. Based on theoretical concepts (eg, Frank-Starling law and dermatome map), examinees are asked to filter the question stem for relevant patient data and adapt the underlying theory to the patient case. This novel insight into ChatGPT points toward persistent deficits in abstract thinking. Therefore, test question writers for the USMLE or other medical examinations may use this question style for other subjects to reduce the risk of AI cheating. Further, our analysis demonstrated that the performance of ChatGPT 4 significantly correlated ($\rho=-0.344; P<0.001$) with the level of test question difficulty. This indicates that sophisticated USMLE questions still challenge and fool both human examinees and AI chatbots. Typically, the most difficult USMLE questions include distractors as well as irrelevant or additional information.; they also require high-level reasoning and interdisciplinary thinking. Our group previously showed that ChatGPT 3.5, similar to the human user peer group, struggled to answer 4- and 5-hammer questions [11]. Such pitfalls continue to perplex the next generation of AI-powered chatbots. Therefore, a thorough analysis of 4- or 5-hammer questions may help examiners refine their test questions and shield the USMLE against AI cheating.

Overall, the phenomenal improvement in the test-taking performance of ChatGPT 4 compared to ChatGPT 3.5 raises intriguing questions regarding future applications and implications of AI in medical education and diagnostics. AI has shown its prowess not only on the USMLE examinations in medical education but also on advanced examinations, such as the neurosurgical written boards [16]. This phenomenon ventures into other aspects of medicine as well, including research and clinical performance [17]. It is imperative that future research ventures into a deeper analysis of the performance of ChatGPT 4 by conducting thorough investigations that probe its strengths and limitations in a more granulated manner, potentially employing diversified medical question banks, simulating real-world scenarios, and engaging experts for analysis and evaluation to allow for the best possible medical education and ultimately patient health care [18].

**Limitations**

This study needs to be interpreted in the light of the following limitations: first, due to the restricted use of ChatGPT (only 25 entries every 3 hours) we were not able to perform a direct comparison of ChatGPT 3.5 and ChatGPT 4 for all test questions included in this study, which might limit its validity. Furthermore, although we attempted to ensure that the questions provided for analysis were not freely available on the internet to minimize the risk of ChatGPT having already seen the exact question, students and researchers around the world may have input certain AMBOSS USMLE Step 3 Style Questions into ChatGPT. This adds a potential confounding factor of ChatGPT memorizing the correct answer from seeing the question beforehand. We used the 2 most recent versions of ChatGPT (ie, ChatGPT 3.5 and ChatGPT 4) to test and compare the performance of large language models on 1840 AMBOSS USMLE Step 3 questions. Thus, the findings of this study should be revalidated for upcoming ChatGPT versions. Future studies may involve additional chatbots, question banks, and image-based test questions. Further, the performance of ChatGPT on USMLE steps could be compared to other national medical licensing exams.

**Conclusions**

This study is the first direct comparison of ChatGPT 4 and ChatGPT 3.5 based on 1840 AMBOSS USMLE Step 3 test questions. Our analysis showed that ChatGPT 4 outperformed its predecessor version across different specialties and difficulty levels, ultimately yielding accuracy levels of 84.7%. However, we could identify persisting weak points of ChatGPT 4, including abstract thinking and elaborated test questions. This line of research may serve as an evidence-based fundament to safeguard the USMLE steps and medical education against AI cheating while underscoring the potency of AI-driven chatbots.

**Conflicts of Interest**

None declared.

**References**


13. AMBOSS question bank. URL: https://www.amboss.com/us [accessed 2023-12-18]


Abbreviations

AI: artificial intelligence
CK: clinical knowledge
CS: communication skills
MRI: magnetic resonance imaging
USMLE: United States Medical Licensing Examination

©Leonard Knoedler, Michael Allertshofer, Samuel Knoedler, Cosima C Hoch, Paul F Funk, Sebastian Cotofana, Bhagvat Maheta, Konstantin Frank, Vanessa Brébant, Lukas Prantl, Philipp Lamby. Originally published in JMI Med Education
Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects

Lukas Weidener¹, BSc, Dr med; Michael Fischer¹, PhD
Research Unit for Quality and Ethics in Health Care, UMIT TIROL – Private University for Health Sciences and Health Technology, Hall in Tirol, Austria

Corresponding Author:
Lukas Weidener, BSc, Dr med
Research Unit for Quality and Ethics in Health Care
UMIT TIROL – Private University for Health Sciences and Health Technology
Eduard-Wallnöfer-Zentrum 1
Hall in Tirol, 6060 Austria
Phone: 43 17670491594
Email: lukas.weidener@edu.umit-tirol.at

Abstract

Background: The use of artificial intelligence (AI) in medicine not only directly impacts the medical profession but is also increasingly associated with various potential ethical aspects. In addition, the expanding use of AI and AI-based applications such as ChatGPT demands a corresponding shift in medical education to adequately prepare future practitioners for the effective use of these tools and address the associated ethical challenges they present.

Objective: This study aims to explore how medical students from Germany, Austria, and Switzerland perceive the use of AI in medicine and the teaching of AI and AI ethics in medical education in accordance with their use of AI-based chat applications, such as ChatGPT.

Methods: This cross-sectional study, conducted from June 15 to July 15, 2023, surveyed medical students across Germany, Austria, and Switzerland using a web-based survey. This study aimed to assess students’ perceptions of AI in medicine and the integration of AI and AI ethics into medical education. The survey, which included 53 items across 6 sections, was developed and pretested. Data analysis used descriptive statistics (median, mode, IQR, total number, and percentages) and either the chi-square or Mann-Whitney U tests, as appropriate.

Results: Surveying 487 medical students across Germany, Austria, and Switzerland revealed limited formal education on AI or AI ethics within medical curricula, although 38.8% (189/487) had prior experience with AI-based chat applications, such as ChatGPT. Despite varied prior exposures, 71.7% (349/487) anticipated a positive impact of AI on medicine. There was widespread consensus (385/487, 74.9%) on the need for AI and AI ethics instruction in medical education, although the current offerings were deemed inadequate. Regarding the AI ethics education content, all proposed topics were rated as highly relevant.

Conclusions: This study revealed a pronounced discrepancy between the use of AI-based (chat) applications, such as ChatGPT, among medical students in Germany, Austria, and Switzerland and the teaching of AI in medical education. To adequately prepare future medical professionals, there is an urgent need to integrate the teaching of AI and AI ethics into the medical curricula.

(JMIR Med Educ 2024;10:e51247) doi:10.2196/51247

KEYWORDS
artificial intelligence; AI technology; medicine; medical education; medical curriculum; medical school; AI ethics; ethics

Introduction

Background
Artificial intelligence (AI) has attracted both public and scientific interest and is amplified by the emergence and greater accessibility of chat-based applications such as ChatGPT (OpenAI, LLC) and Bard (Google, LLC). For several years, the medical field has been an active and expanding area of research on the application of AI [1]. As of now, AI is used in diverse medical specializations, including dermatology, radiology, and pathology [2-4].
Although the history of AI can be traced back to the 1950s, the public’s unrestricted access to highly advanced large language models, such as ChatGPT, can be seen as a significant turning point in the history of AI [5,6]. Early studies demonstrated that ChatGPT is capable of successfully completing the written portion of the United States Medical Licensing Examination [7]. Given the capabilities of AI-based chat applications such as ChatGPT in medicine, further studies have highlighted their potential use in providing information on cancer, assisting in clinical diagnoses, authoring scientific research articles, and patient communication [8-10]. Considering the wide availability and integration of medical knowledge in this application, its increasing use in medicine and among medical students is foreseeable [11].

Despite the long history of AI and the increasing adoption of this technology, there is disagreement regarding its definition among the scientific community [12]. There is a consensus within the scientific community on distinguishing between the so-called strong AI, also known as “artificial general intelligence,” and weak AI or “artificial narrow intelligence” [13]. This categorization is based on the capabilities of AI or its areas of application [13]. Strong AI, recognized for its human-equivalent intellectual abilities and knowledge, stands in contrast to weak AI, which refers to AI solutions capable of accomplishing specific tasks effectively [13]. The area of weak AI can be further divided into the so-called statistical AI and symbolic AI. The field of statistical AI also includes machine learning and deep learning, on which large language models such as ChatGPT are based [13]. Areas of application for symbolic AI in medicine include expert systems (eg, clinical decision support systems), which make decisions based on explicit knowledge in the form of predefined rules [14].

Considering the likely significant impact the implementation and use of AI in medicine is poised to make, a growing body of literature advocates the inclusion of AI-related content in medical curricula [15-18]. In addition to implications for the medical profession and patient care, medical students are expected to face new ethical challenges posed by the use of AI in medicine [15,19]. Despite the potentially significant ethical challenges anticipated from the deployment of AI in medicine, such as the possibility of discrimination due to biases in the data used for training or effects on patient autonomy, there is a near-complete absence of scientific publications on specific teaching content or methods related to AI ethics as part of medical higher education.

In addition to the lack of specificity regarding teaching content on AI and AI ethics, the absence of studies on medical students’ perception of AI ethics education (including teaching content) is notable [20,21]. It is essential to point out that the current state of research regarding medical students’ perceptions and assessments of AI application in medicine largely represents a knowledge base that predates the advent of large language models such as ChatGPT. With the ubiquity of the aforementioned AI applications at the time of this publication, it is reasonable to expect that medical students’ assessments of AI implementation in medicine will deviate significantly from earlier publications within this area of research, highlighting the need for further research.

Objective
This study aimed to explore how medical students perceive the use of AI in medicine, as well as the teaching of AI and AI ethics (including prospective AI ethics teaching topics). In this context, the introduction and accessibility of large language models such as ChatGPT should be emphasized, leading to the following research question: how do medical students from Germany, Austria, and Switzerland perceive (1) the application of AI in medical practice, (2) the integration of AI and AI ethics into medical education, and (3) AI ethics teaching content in their curriculum in accordance with the use of AI-based chat applications such as ChatGPT?

To address this research question, the participating medical students were divided into 2 groups based on their prior use of AI-based (chat) applications, such as ChatGPT.

Methods
Overview
This cross-sectional study was conducted between June 15 and July 15, 2023. During this time frame, an invitation to participate in the study was sent to medical students who were regularly enrolled in universities in Germany, Austria, and Switzerland. The study sample included medical students from all academic semesters, including those in practically oriented semesters such as the practical year in Germany. Participation in the study was voluntary and there were no consequences for nonparticipation. The study used an anonymous web-based survey, with recruitment facilitated through email invitations and assistance from various medical student associations, unions, and councils in their respective countries. To minimize potential selection bias, the survey invited medical students from various universities and academic semesters in Germany, Austria, and Switzerland. This strategy ensured a broad and representative sample of the participants. Moreover, careful construction and pretesting of the survey were conducted to minimize potential response biases. Before the official data collection, a pretest was conducted with 11 medical students from the target population. The web-based survey provider, “LimeSurvey” was used for both the pretest and the main study.

Ethical Considerations
The Research Committee for Scientific Ethical Questions granted ethical approval for this study (3181) on January 16, 2023.

Survey Development
The survey used for data collection was developed based on existing scientific publications [15,22]. Owing to the lack of references in the areas of AI teaching, AI ethics, and recent developments in AI, most items used for the survey were newly formulated. The survey comprises 53 items, including both questions and statements. During the development process, these items were distributed across 6 parts, with some contingent on the responses to the preceding items. The first part aimed to collect information on the demographic characteristics and educational background of the participants. To address the research question of this study, participants were divided into...
2 groups based on their responses to questions related to their prior use of AI-based (chat) applications such as ChatGPT. The second part sought to gather information about the students’ previous experiences with AI-based (chat) applications. In the third part, the students were asked to rate various statements regarding the use of AI in medicine. The fourth and fifth parts aimed to capture students’ evaluations of statements about AI teaching and ethics, respectively. The sixth part assessed the perceived relevance of the potential teaching content to AI ethics. The items in parts 3 to 6 were evaluated using a 5-point Likert scale. Before the survey was conducted, 2 experts in ethics and AI evaluated the survey and their recommendations were incorporated. Upon receiving expert feedback, the teaching topic of “data privacy” was introduced as a distinct subject under AI ethics. Previously, this was encompassed within the broader “safety” category. Furthermore, to enhance clarity, the term “knowingly” was incorporated into Q12. This adjustment acknowledges that the application of AI in medicine may not always be transparent.

Survey Pretest

To assess the comprehensibility and relevance of the survey, a pretest was conducted with 11 medical students, who subsequently provided feedback. This feedback led to 6 relevant modifications aimed at enhancing clarity, relevance, and user-friendliness. Because of the feedback provided, questions Q1 through Q4 and Q6 were specified by adding examples following each question. The changes made to the questions are highlighted in italics:

1. Q1. Have you already received education in the field of ethics within your regular medical studies? (eg, as part of the History, Ethics, and Theory of Medicine course)
2. Q2. Have you already received education in the field of AI in your regular medical studies? (eg, as part of medical statistics or informatics)
3. Q3. Have you already received education in the field of AI outside of your regular medical studies? (eg, in the form of further training, own research)
4. Q4. Have you already received education in the field of AI ethics within your regular medical studies? (eg, as part of the History, Ethics, and Theory of Medicine course)
5. Q6. Have you already received instruction in the field of AI ethics outside of your regular medical studies? (eg, in the form of further training, own research)

Similarly, statement 27 (S27) was further improved by adding examples from various fields to underscore the multidisciplinary context: “AI ethics should be taught by experts from various fields (eg, medicine, computer science, philosophy) to ensure a multidisciplinary perspective on AI ethics.”

To improve the survey’s user experience, conditional logic was integrated so that questions Q5 and Q7 appeared only in response to the specific preceding answers. Both question Q5 and question Q7 were designed to explore the specific content covered in AI ethics education. These questions were identical in wording: “Which of the following contents were covered as part of the instruction/education?” Question Q5 was presented exclusively to participants who answered “yes” to question 4, focusing on AI ethics education outside of their regular medical curriculum. This strategic modification not only streamlined the survey’s presentation but also minimized the immediate visual content, reducing complexity.

Sample Size Calculation

The sample size for this study was calculated before data collection using Cochran sample size formula \( n = \left[ Z^2 * p * (1-p) / E^2 \right] \). The total population size used for the calculation, which represents the number of medical students enrolled at the end of the winter semester in 2022, was 130,601 across the 3 countries included in the study. This figure includes 105,275 medical students from Germany (accounting for 80.61% of the total), 17,826 from Austria (13.65%), and 7500 from Switzerland (5.74%) [24-26]. This summation was performed based on the primary research question and was predicated on the assumption that the prevalence of AI-based (chat) applications, such as ChatGPT, among medical students does not vary significantly across these countries. A confidence level of 95% \( (Z=1.96) \) and a margin of error of 5% were used to determine the sample size. The proportion \( (p) \) was derived from a pretest involving a separate group of 11 medical students of which 5 were already using large language models such as ChatGPT before the study \( (P=.45) \). Cochran’s formula yielded a sample size of 380 medical students. As the study was conducted using a web-based survey with recruitment via email, an estimated dropout rate of 40% was factored in. To achieve a calculated sample size of 380 participants, at least 532 students were targeted during the recruitment process. To ensure adequate representation based on the proportion of medical students within each country of interest, the study aimed to include at least 306 medical students from Germany, 52 from Austria, and 22 from Switzerland in the data collection and analysis process.

Note that these are rounded values given that the actual calculations result in noninteger numbers.

Data Analysis

Collected data were evaluated using SPSS (version 28; IBM Corp), LimeSurvey (LimeSurvey GmbH), and Microsoft Excel (version 16.73). Descriptive statistics were calculated for all survey variables, including the median, IQR, mode, total number, and percentages. For further statistical analysis, the chi-square test of independence was used to compare the 3 groups. When significant differences were observed in the chi-square test, post hoc analysis was performed using the adjusted residuals method to specify which specific groups or categories contributed to the observed significance. In addition, \( z \) scores were calculated to facilitate the comparison of responses across different groups. These were computed using the 2-sided test formula \( z = (X - \mu) / \sigma \), where \( X \) represents the value of the response, \( \mu \) is the mean of the responses for the group, and \( \sigma \) is the SD within that group. The calculation of \( z \) scores enabled the quantification of the deviation of each response from the group mean in terms of SDs. The Mann-Whitney \( U \) test was used for the statistical comparison of 2 independent groups; for further statistical analysis, the chi-square test of independence was used to compare the 3 groups, and the Mann-Whitney \( U \)
test was used for the statistical comparison of 2 independent groups. For statistical analysis, the responses to the Likert scale were recoded into a numerical format (“I strongly disagree”=1, “I disagree”=2, “undecided”=3, “I agree”=4, “I strongly agree”=5). For all statistical tests performed, the significance level was set at $\alpha=.05$, and a value of $P\leq .05$ was considered statistically significant. Only complete data sets were included in the data analysis to avoid potential biases that could arise from replacing or estimating the missing values (list-wise deletion).

**Results**

**Overview**

In total, 521 medical students participated in the survey, yielding 487 complete and valid data sets for the statistical analysis. The survey invitations were disseminated via email with the help of medical student associations, unions, and councils. The total number of medical students reached and the precise response rate could only be approximated. On the basis of the feedback received from the engaged medical student councils, we estimated that at least 2000 medical students were approached. This would be equal to a response rate of 24.35% (487/2000). Our sample size calculation was based on the assumption that the use of AI-based (chat) applications such as ChatGPT does not diverge markedly among medical students from each of the countries of interest, namely Germany, Austria, and Switzerland. Consequently, the chi-square test of independence was used for statistical evaluation. We posited a null hypothesis (H) asserting no association between the variables (use of AI-based applications and country of study) and an alternative hypothesis (H₁) suggesting an association between these variables. The chi-square test returned a value of $P=.96$, which exceeded the predetermined level of significance. As such, we did not reject the null hypothesis, leading us to conclude that there is no statistically significant association between the use of AI-based (chat) applications and country of study among the surveyed medical students, given that each individual fits into one category for each variable.

**Part 1: Demographics and Educational Background**

Of the medical students who participated in the survey, the majority were women (270/487, 55.4%). The largest demographic age was between 20 and 25 years (301/487, 61.8%), and most students were enrolled in Germany (296/487, 60.7%). The German contingent of respondents was slightly below our target size of 306, representing a 3.3% (296/306) shortfall. However, participation from Austria exceeded our initial target of 52 students by a substantial margin, with 105 respondents indicating enrollment in Austria, denoting an overachievement rate of 202% (105/52). Similarly, Swiss representation surpassed our initial target of 22 students, with 86 respondents registered in Switzerland, marking an overachievement of 391% (86/22). Most of the surveyed students were in the clinical stage (CS) of their study (277/487, 56.9%), followed by those in their practical years (63/487, 12.9%). Comprehensive demographic characteristics are presented in Table 1.

The respondents were also asked about their educational backgrounds in ethics, AI, and AI ethics. Most participants (425/487, 87.2%) reported having received ethics education. However, a considerably smaller proportion of respondents claimed that they had received prior education in AI as part of their medical curriculum (26/487, 5.3%), with an additional 10.5% (51/487) having obtained such knowledge outside of their regular medical studies. Few participants had been exposed to AI ethics education within their medical curriculum (21/487, 4.3%), with a small number reporting having learned about AI ethics outside their regular curriculum (51/487, 6.8%). The most common subjects covered in AI ethics education were bias (15/487, 3.1% within and 14/487, 2.9% outside regular studies) and explainability (12/487, 2.5% within and 20/487, 4.1% outside regular studies). Detailed responses related to the participants’ educational background are shown in Table 2.
<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Medical students, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>270 (55.4)</td>
</tr>
<tr>
<td>Man</td>
<td>203 (41.7)</td>
</tr>
<tr>
<td>Nonbinary</td>
<td>3 (0.6)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>11 (2.3)</td>
</tr>
<tr>
<td><strong>Age (y)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;20</td>
<td>56 (11.5)</td>
</tr>
<tr>
<td>20-25</td>
<td>301 (61.8)</td>
</tr>
<tr>
<td>26-30</td>
<td>92 (18.9)</td>
</tr>
<tr>
<td>31-35</td>
<td>28 (5.7)</td>
</tr>
<tr>
<td>&gt;35</td>
<td>10 (2.0)</td>
</tr>
<tr>
<td><strong>Country of enrollment (medical studies)</strong></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>296 (60.7)</td>
</tr>
<tr>
<td>Austria</td>
<td>105 (21.5)</td>
</tr>
<tr>
<td>Switzerland</td>
<td>86 (17.7)</td>
</tr>
<tr>
<td><strong>Stage of study</strong></td>
<td></td>
</tr>
<tr>
<td>Preclinical</td>
<td>57 (11.7)</td>
</tr>
<tr>
<td>Clinical</td>
<td>277 (56.9)</td>
</tr>
<tr>
<td>Practical year</td>
<td>63 (12.9)</td>
</tr>
<tr>
<td>Elective year</td>
<td>26 (5.3)</td>
</tr>
<tr>
<td>Bachelor</td>
<td>46 (9.4)</td>
</tr>
<tr>
<td>Master</td>
<td>18 (3.7)</td>
</tr>
</tbody>
</table>
Table 2. Educational background of the participating medical students from Germany, Austria, and Switzerland (n=487).

<table>
<thead>
<tr>
<th>Question</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1: Have you already received education in the field of ethics within your regular medical studies? (eg, as part of the History, Ethics, and Theory of Medicine course)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>425 (87.2)</td>
</tr>
<tr>
<td>No</td>
<td>62 (12.7)</td>
</tr>
<tr>
<td>Q2: Have you already received education in the field of artificial intelligence within your regular medical studies? (eg, as part of medical statistics or informatics)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>26 (5.3)</td>
</tr>
<tr>
<td>No</td>
<td>461 (94.7)</td>
</tr>
<tr>
<td>Q3: Have you already received education in the field of artificial intelligence outside of your regular medical studies? (eg, in the form of further training, own research)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>51 (10.5)</td>
</tr>
<tr>
<td>No</td>
<td>436 (89.2)</td>
</tr>
<tr>
<td>Q4: Have you already received education in the field of artificial intelligence ethics within your regular medical studies? (eg, as part of the History, Ethics, and Theory of Medicine course)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>21 (4.3)</td>
</tr>
<tr>
<td>No</td>
<td>466 (95.7)</td>
</tr>
<tr>
<td>Q5: Which of the following contents were covered as part of the education?&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Informed consent</td>
<td>11 (2.3)</td>
</tr>
<tr>
<td>Bias</td>
<td>15 (3.1)</td>
</tr>
<tr>
<td>Data privacy</td>
<td>13 (2.7)</td>
</tr>
<tr>
<td>Explainability</td>
<td>12 (2.5)</td>
</tr>
<tr>
<td>Safety (of AI-based applications)</td>
<td>10 (2)</td>
</tr>
<tr>
<td>Fairness</td>
<td>5 (1)</td>
</tr>
<tr>
<td>Autonomy</td>
<td>8 (1.6)</td>
</tr>
<tr>
<td>Responsibility</td>
<td>8 (1.6)</td>
</tr>
<tr>
<td>Q6: Have you already received education in the field of artificial intelligence ethics outside of your regular medical studies? (eg, in the form of further training, own research)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>33 (6.8)</td>
</tr>
<tr>
<td>No</td>
<td>454 (93.2)</td>
</tr>
<tr>
<td>Q7: Which of the following contents were covered as part of the education?&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Informed consent</td>
<td>10 (2)</td>
</tr>
<tr>
<td>Bias</td>
<td>14 (2.9)</td>
</tr>
<tr>
<td>Data privacy</td>
<td>17 (3.5)</td>
</tr>
<tr>
<td>Explainability</td>
<td>20 (4.1)</td>
</tr>
<tr>
<td>Safety (of artificial intelligence-based applications)</td>
<td>18 (3.7)</td>
</tr>
<tr>
<td>Fairness</td>
<td>12 (2.5)</td>
</tr>
<tr>
<td>Autonomy</td>
<td>14 (2.9)</td>
</tr>
<tr>
<td>Responsibility</td>
<td>19 (3.9)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Question 5 was exclusively displayed to participants who responded to question 4 with “yes.”

<sup>b</sup>An explanation of the contents of Q5 and Q7 is provided in the text.

<sup>c</sup>Question 7 was exclusively displayed to participants who responded to question 6 with “yes.”
Part 2: Use of AI-Based (Chat) Applications

With regard to the use of AI-based (chat) applications such as ChatGPT (OpenAI), Bard (Google), Bing Chat (Microsoft Inc), and Jasper Chat (Jasper AI, Inc), 38.8% (189/487) of the respondents reported prior use of these platforms. Conversely, the vast majority (438/487, 89.9%) indicated that they did not knowingly use other AI-based medical applications. Of the 298 respondents who had not previously used an AI-based chat application, 76.9% (n=229) expressed an interest in future use. Among the respondents who reported prior use of AI-based (chat) applications, nearly half had used such an application for 1-3 hours over the past week (91/189, 48.2%). Of this group, 73% (138/189) indicated using an AI-based (chat) application in a medical context, with the most common use being querying medical knowledge (74/138, 53.6%). The results of this survey are summarized in Table 3.

Table 3. Answers to the use of AI-based (chat) applications of participants (n=487).

<table>
<thead>
<tr>
<th>Question</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q8: Have you already used an AI-based (chat) application such as ChatGPT (OpenAI), Bard (Google), Bing chat, or Jasper Chat?</td>
<td>189 (38.8)</td>
</tr>
<tr>
<td>Yes</td>
<td>189 (38.8)</td>
</tr>
<tr>
<td>No</td>
<td>298 (61.2)</td>
</tr>
<tr>
<td>Q9: Have you knowingly ever used AI-based medical applications, such as image-based diagnostic tools in radiology?</td>
<td>49 (10.1)</td>
</tr>
<tr>
<td>Yes</td>
<td>49 (10.1)</td>
</tr>
<tr>
<td>No</td>
<td>438 (89.9)</td>
</tr>
<tr>
<td>Q10: Are you interested in using an AI application as part of your medical studies in the future?^b; n=298</td>
<td>229 (76.9)</td>
</tr>
<tr>
<td>Yes</td>
<td>229 (76.9)</td>
</tr>
<tr>
<td>No</td>
<td>69 (23.1)</td>
</tr>
<tr>
<td>Q1: Approximately how many hours have you used the AI-based (chat) application in the last week (7 d)^c; (n=189)</td>
<td>73 (38.6)</td>
</tr>
<tr>
<td>&lt;1 h</td>
<td>73 (38.6)</td>
</tr>
<tr>
<td>1-3 h</td>
<td>91 (48.2)</td>
</tr>
<tr>
<td>4-6 h</td>
<td>19 (10)</td>
</tr>
<tr>
<td>7-9 h</td>
<td>3 (1.6)</td>
</tr>
<tr>
<td>10-12 h</td>
<td>2 (1.1)</td>
</tr>
<tr>
<td>&gt;12 h</td>
<td>1 (0.5)</td>
</tr>
<tr>
<td>Q12: Have you already used the AI-based (chat) application in a medical context? (eg, for explaining medical conditions or medical questions)^d; (n=189)</td>
<td>138 (73)</td>
</tr>
<tr>
<td>Yes</td>
<td>138 (73)</td>
</tr>
<tr>
<td>No</td>
<td>51 (26.7)</td>
</tr>
<tr>
<td>Q13: For which of the following objectives have you already used the AI-based (chat) application in the medical context?^e; (n=138)</td>
<td>18 (13)</td>
</tr>
<tr>
<td>Therapy suggestions</td>
<td>18 (13)</td>
</tr>
<tr>
<td>Querying medical knowledge</td>
<td>74 (53.6)</td>
</tr>
<tr>
<td>Diagnostic support</td>
<td>5 (3.6)</td>
</tr>
<tr>
<td>Explanation of pathologies</td>
<td>41 (29.7)</td>
</tr>
</tbody>
</table>

^aAI: artificial intelligence.
^bQuestion 10 was exclusively displayed to participants who responded to questions 8 and 9 with “no.”
^cQuestion 11 was exclusively displayed to participants who responded to question 8 with “yes.”
^dQuestion 12 was exclusively displayed to participants who responded to question 8 with “yes.”
^eQuestion 13 was exclusively displayed to participants who responded to question 12 with “yes.”

Part 3: AI in Medicine

In the third part of the survey, participants’ attitudes toward the role of AI’s in medicine were examined. Of the 487 respondents, 71.7% (n=349) agreed or strongly agreed that the use of AI would bring about positive changes to medicine (S1). Similarly, 72.1% (350/487) believed that AI could find practical applications in medicine (S2). When comparing the responses between those who had used AI-based applications and those who did not, significant differences were identified for each statement using the Mann-Whitney U test (S1: P=.003; S2: P=.002). Although both groups shared the same median and mode responses, their z scores suggested variations in their agreement levels. Specifically, respondents who had not
previously used AI-based chat applications displayed a higher level of agreement with the statement in S1 (z score: -2.991). Conversely, those who had used AI-based applications exhibited greater concurrence with the statement in S2 (z score: 3.105).

When comparing the responses of those who had used AI-based chat applications and those who had not, no significant difference was observed regarding the subsequent 2 statements, S3 and S4, which were related to the influence on the choice of medical specialization and the potential reduction of jobs for medical staff. However, marked differences were identified when comparing the responses to statements S5 to S7 concerning improvements in patient care quality (S5: \( P < .001 \)), diagnostic processes (S6: \( P = .002 \)), and therapy selection (S7: \( P < .001 \)). Although the overall agreement (either “agree” or “strongly agree”) was high for these statements (S5: 71%; S6: 76.4%; S7: 77.9%), z scores indicated greater agreement within the subgroup that had previously used AI-based (chat) applications (S5: \( z \) score=3.570; S6: \( z \) score=3.089; S7: \( z \) score=3.865).

No significant difference was found for statements S8 to S11 between the 2 groups, with comparable levels of overall agreement (“agree” or “strongly agree”) for each statement (S8: 31.8%; S9: 29.6%; S10: 25.9%; S11: 31.8%). However, a significant difference was observed for statement S12 (\( P = .02 \)), with 95.3% of all respondents agreeing or strongly agreeing that the use of AI in medicine presents new ethical challenges. The \( z \) score (2.302), median (5), and mode (5) suggested a higher level of agreement among the groups that had previously used AI-based (chat) applications, such as ChatGPT. A statistical analysis of the third part of the survey is presented in Table 4.

A detailed illustration of the perceptions of the surveyed medical students regarding the use of AI in medicine is provided in Table S1 in Multimedia Appendix 1.
Table 4. Statistical analysis of the perceptions of medical students regarding the use of artificial intelligence (AI)–based (chat) applications such as ChatGPT (OpenAI), Bard (Google), Bing Chat (Microsoft Inc), and Jasper Chat (Jasper AI, Inc) in medicine (n=487).

<table>
<thead>
<tr>
<th>Statement and subgroup</th>
<th>Scores, median (IQR)</th>
<th>Scores, mode</th>
<th>P value</th>
<th>Z score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The use of AI in medicine will...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S1: ...positively change medicine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3.75-4.25)</td>
<td>4</td>
<td>.003</td>
<td>-2.990</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S2: ...find useful applications in medicine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.002</td>
<td>3.101</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S3: ...influence the choice of my medical specialization</strong></td>
<td></td>
<td></td>
<td>.52</td>
<td>-1.474</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>3 (2-4)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>3 (2-4)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S4: ...reduce the number of jobs for medical staff</strong></td>
<td></td>
<td></td>
<td>.09</td>
<td>-1.707</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>3 (3-5)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>3 (2-4)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S5: ...improve the quality of patient care</strong></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>3.570</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (0)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S6: ...improve the process of diagnosis</strong></td>
<td></td>
<td></td>
<td>.002</td>
<td>3.089</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S7: ...improve the process of therapy selection</strong></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>3.865</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (0-0)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S8: ...negatively affect the doctor-patient relationship</strong></td>
<td></td>
<td></td>
<td>.18</td>
<td>1.328</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>3 (2-4)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>3 (2-4)</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S9: ...lead to a dehumanization of medicine</strong></td>
<td></td>
<td></td>
<td>.11</td>
<td>1.610</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>3 (2-4)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>3 (2-4)</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S10: ...negatively affect patient autonomy</strong></td>
<td></td>
<td></td>
<td>.05</td>
<td>2.040</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>3 (2-3)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>3 (2-4)</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S11: ...negatively affect the autonomy of medical staff</strong></td>
<td></td>
<td></td>
<td>.16</td>
<td>1.415</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>3 (2-4)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>3 (2-4)</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S12: ...bring new ethical challenges</strong></td>
<td></td>
<td></td>
<td>.02</td>
<td>2.302</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>5 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part 4: Teaching AI in Medical Education

When asked about their agreement on whether AI teaching should be incorporated into medical education (S13), 74.9% (385/487) of the respondents agreed or strongly agreed. A statistically significant difference was identified between those with and without prior use of AI-based (chat) applications ($P=.02$). The mean (5), mode (5), and $z$ score (2.381) suggest higher agreement within the group that previously used AI-based applications. In contrast, there was an overall disagreement
(88%) with the assertion that AI instruction in medical education is currently sufficient (S14), with no statistically significant difference between the 2 groups. No significant statistical differences were observed for statements S15-S19. There was an overall agreement that the teaching of AI should include practical content (S15; 417/487, 86%), be based on case studies and application scenarios in medicine (S16; 342/487, 70.3%), be an important prerequisite for medical practice (S17; 314/487, 64.9%), be available to medical staff even after graduation (S18; 376/487, 77.3%), and be updated regularly to reflect advances in AI technology (S19; 407/487, 83.6%). There was a significant measurable difference in the S20 ($P = .002$) between the 2 groups. The $z$ score indicates a stronger agreement with the statement “AI instruction is of interest to me” among the group of medical students who previously used AI-based (chat) applications ($z$ score: 3.173). The statistical analysis is presented in Table 5, and an overview of the perceptions of the participants regarding the teaching of AI in medicine can be found in Table S2 in Multimedia Appendix 1.

### Table 5. Statistical analysis of the perceptions of medical students regarding the teaching of artificial intelligence (AI)-based (chat) applications such as ChatGPT (OpenAI), Bard (Google), Bing Chat (Microsoft Inc), and Jasper Chat (Jasper AI, Inc) in medical education (n=487).

<table>
<thead>
<tr>
<th>Statement and subgroup</th>
<th>Scores, median (IQR)</th>
<th>Scores, mode</th>
<th>$P$ value</th>
<th>$Z$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The teaching of AI...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S13: ...should be part of medical education</td>
<td>5 (4-5)</td>
<td>5</td>
<td>.02</td>
<td>2.381</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>5 (4-5)</td>
<td>5</td>
<td>.02</td>
<td>2.381</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td>.02</td>
<td>2.381</td>
</tr>
<tr>
<td>S14: ...in medical education is adequate</td>
<td>2 (1-2)</td>
<td>1</td>
<td>.90</td>
<td>0.128</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>2 (1-2)</td>
<td>1</td>
<td>.90</td>
<td>0.128</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>2 (1-2)</td>
<td>1</td>
<td>.90</td>
<td>0.128</td>
</tr>
<tr>
<td>S15: ...should include practical content (e.g., exercises to apply AI) in addition to theoretical aspects</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.18</td>
<td>−2.358</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.18</td>
<td>−2.358</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (0)</td>
<td>4</td>
<td>.18</td>
<td>−2.358</td>
</tr>
<tr>
<td>S16: ...should be based on case studies and application scenarios of AI in medicine</td>
<td>4 (3-5)</td>
<td>4</td>
<td>.53</td>
<td>−0.625</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3-5)</td>
<td>4</td>
<td>.53</td>
<td>−0.625</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.53</td>
<td>−0.625</td>
</tr>
<tr>
<td>S17: ...is an important prerequisite for medical practice</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.16</td>
<td>1.417</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.16</td>
<td>1.417</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td>.16</td>
<td>1.417</td>
</tr>
<tr>
<td>S18: ...should be available for medical staff even after graduation</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.13</td>
<td>−1.527</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td>.13</td>
<td>−1.527</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-5)</td>
<td>4</td>
<td>.13</td>
<td>−1.527</td>
</tr>
<tr>
<td>S19: ...should be updated regularly to reflect advances in AI technology</td>
<td>4 (3-4)</td>
<td>4</td>
<td>.34</td>
<td>−2.121</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td>.34</td>
<td>−2.121</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td>.34</td>
<td>−2.121</td>
</tr>
<tr>
<td>S20: ...is of interest to me</td>
<td>4 (4-5)</td>
<td>4</td>
<td>.002</td>
<td>3.173</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>4</td>
<td>.002</td>
<td>3.173</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (0)</td>
<td>4</td>
<td>.002</td>
<td>3.173</td>
</tr>
</tbody>
</table>

### Part 5: Teaching AI Ethics in Medical Education

In the survey, 74.9% (385/487) of medical students agreed or strongly agreed that teaching AI ethics should be included in medical education (S21). However, only 4.9% (24/487) agreed that the current instruction on AI ethics in medical education is adequate (S22). For statements S23 to S27, the vast majority of medical students generally agreed ("agree" or "strongly agree") that the teaching of AI ethics should be based on case studies and application scenarios of AI in medicine (S23; 412/487, 85%), contribute to raising awareness of ethical issues in medical practice (S24; 343/487, 70.6%), is an important prerequisite for medical practice (S25; 354/487, 72.8%), should be available for medical staff even after graduation (S26; 370/487, 75.9%), and should be taught by experts from various fields (eg, medicine, computer science, and philosophy) to ensure a multidisciplinary perspective on AI ethics (S27; 416/487, 85.2%). No statistically significant differences were observed for statements S21 to S27 between the 2 groups (those with previous use of AI-based [chat] applications and those without). Despite the $z$ score of 1.782 being below the typical...
threshold of 1.96 for a 2-tailed test, the statement “the teaching of AI ethics is of interest to me” (S28) showed a statistically significant difference \( (P = .005) \). This indicates that even though the deviation from the mean agreement level is not as strong as typically expected for significance, those who had previously used AI-based (chat) applications demonstrated a notably higher level of interest in AI ethics teaching than those who had not. The statistical analysis for part 5 of the survey is shown in Table 6, and the distribution of answers is presented in Table S3 in Multimedia Appendix 1.

Table 6. Statistical analysis of the perceptions of medical students regarding the teaching of artificial intelligence (AI)-based (chat) applications such as ChatGPT (OpenAI), Bard (Google), Bing Chat (Microsoft Inc), and Jasper Chat (Jasper AI, Inc) ethics in medical education (n=487).

<table>
<thead>
<tr>
<th>Statement and subgroup</th>
<th>Scores, median (IQR)</th>
<th>Scores, mode</th>
<th>( P ) value</th>
<th>( Z ) score</th>
</tr>
</thead>
<tbody>
<tr>
<td>S13: ... should be part of medical education</td>
<td></td>
<td></td>
<td>.37</td>
<td>−0.903</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>5 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S14: ... in medical education is adequate</td>
<td></td>
<td></td>
<td>.21</td>
<td>−1.263</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>2 (2-3)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>2 (1-2)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S15: ... should include practical content (e.g., exercises to apply AI) in addition to theoretical aspects</td>
<td></td>
<td></td>
<td>.80</td>
<td>−0.254</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (0)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (0)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S16: ... should be based on case studies and application scenarios of AI in medicine</td>
<td></td>
<td></td>
<td>.48</td>
<td>−0.707</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (2.5-4.5)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S17: ... is an important prerequisite for medical practice</td>
<td></td>
<td></td>
<td>.90</td>
<td>0.118</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (2-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S18: ... should be available for medical staff even after graduation</td>
<td></td>
<td></td>
<td>.17</td>
<td>−1.359</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (2-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S19: ... should be updated regularly to reflect advances in AI technology</td>
<td></td>
<td></td>
<td>.17</td>
<td>−1.381</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S20: ... is of interest to me</td>
<td></td>
<td></td>
<td>.005</td>
<td>1.782</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (0)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part 6: AI Ethics Teaching Content

In analyzing the perceptions of medical students with and without prior exposure to AI chat applications regarding AI ethics content, all 8 proposed topics were deemed highly relevant (“quite relevant” and “very relevant”) by the respondents: TC1: 418/487, 85.9%; TC2: 408/487, 83.8%; TC3: 384/487, 78.9%; TC4: 415/487, 85.2%; TC5: 423/487, 86.2%; TC6: 407/487, 83.6%; TC7: 402/487, 82.5%; and TC8: 448/487, 92.3%). No statistically significant difference was observed between the responses of both groups, except for TC1 (informed consent; \( P = .04 \)). The \( Z \) score suggests that medical students who had previously used AI-based (chat) applications perceived informed consent to be more relevant than those who had not (\( Z \) score: 2.018). The statistical results of this section are shown in Table 7, with an overview of the statements on the relevance of AI ethics teaching content provided in Table S4 in Multimedia Appendix 1.
### Table 7. Statistical analysis of the relevance of artificial intelligence (AI)–based (chat) applications such as ChatGPT (OpenAI), Bard (Google), Bing Chat (Microsoft Inc), and Jasper Chat (Jasper AI, Inc) ethics teaching contents according to the participating medical students (n=487).

<table>
<thead>
<tr>
<th>AI ethics teaching content and subgroup</th>
<th>Scores, median (IQR)</th>
<th>Scores, mode</th>
<th>P value</th>
<th>Z score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC1: informed consent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td>.04</td>
<td>2.018</td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC2: bias</td>
<td></td>
<td></td>
<td>.22</td>
<td>-1.215</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC3: data privacy</td>
<td></td>
<td></td>
<td>.78</td>
<td>0.283</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC4: explainability</td>
<td></td>
<td></td>
<td>.36</td>
<td>-0.911</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3.5-4.5)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC5: safety</td>
<td></td>
<td></td>
<td>.57</td>
<td>0.565</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>5 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>5 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC6: fairness</td>
<td></td>
<td></td>
<td>.96</td>
<td>-0.048</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC7: autonomy</td>
<td></td>
<td></td>
<td>.11</td>
<td>1.594</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC8: responsibility</td>
<td></td>
<td></td>
<td>.22</td>
<td>-1.215</td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>5 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>5 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Additional Analysis of the Collected Data

To analyze whether there is a difference in education regarding AI and AI ethics among Germany, Austria, and Switzerland, we conducted an additional evaluation of the collected data. For this supplementary analysis, we analyzed the responses to Q2: “Have you already received education in the field of artificial intelligence within your regular medical studies? (eg, as part of medical statistics or informatics),” and Q4: “Have you already received education in the field of AI ethics within your regular medical studies? (eg, as part of the History, Ethics, and Theory of Medicine course).” Using the chi-square test of independence, we sought to determine whether the distribution of answers varied significantly among these countries. In the comparison between the 3 countries concerning education in the field of AI, the chi-square test of independence indicated no significant difference in the distribution of the responses. Of the 487 respondents, only 26 (5.3%) indicated that they had previously received AI education. The test yielded a result of $\chi^2(N=487)=0.1 (P=.33)$. Similarly, regarding education in the field of AI ethics, the distribution of responses among the countries was not significantly different. Of the 487 respondents, only 21 (4.3%) indicated that they had received education on AI ethics. The test yielded a result of $\chi^2(N=487)=0.3 (P=.19).

### Stage of Study

To account for potential confounders, such as the stage of the study, further analyses were performed on the data set. Recognizing the possible overlaps and similarities in experiences and perspectives across the different stages, the original 6 stages of the study were further consolidated. The stages “preclinical” and “bachelor” were summarized into the “preclinical stage (PCS).” Similarly, the “clinical” and “master” stages were combined into the “clinical stage.” Finally, the “practical year” and “elective year” stages were grouped together to form the “clinical practical stage (CPS).” With these redefined categories, the chi-square test of independence was used to analyze whether there were significant variations in perceptions and responses across the 3 consolidated stages.

Focusing on the potential impact of AI in medicine, a significant difference was observed in the statement, “the use of AI in medicine will influence the choice of my specialization” (S3). CPS participants were notably more influenced than those in the PCS ($P=.004$). However, no difference was evident between...
the PCS and CS participants. Most other statements concerning AI’s impact on medicine (S1-2; S4-12) did not demonstrate statistical significance. Similarly, no significant difference was found for statements related to AI teaching (S13-20) across the study stages (PCS, CS, and CPS). When considering the teaching of AI ethics, differences were evident in the belief that AI ethics should be integrated into medical education (S21; \( P=.003 \)) and that the current teaching of AI ethics is adequate (S22; \( P=.02 \)). Upon further analysis, CS participants showed stronger agreement than PCS participants, with no difference when compared with CPS participants. Finally, for the specific content of AI ethics teaching, none of the statements reflected significant statistical variation across the study stages. An overview of the statistical differences is provided in Tables S5-S8 in Multimedia Appendix 1.

**Ethics Education Background**

To explore the potential impact of prior ethics education on survey outcomes, particularly in parts 3 to 6, we compared 2 distinct groups: those with prior ethics education and those without. On the use of AI in medicine, one statistical difference could be determined for the statement that “...negatively affect the autonomy of medical staff” (S11, \( P=.002 \)). The \( z \) score suggested a stronger level of agreement with the statement in the group that had received prior ethics education (\( z \) score: 2.876). For the other statements of the third part of the survey (S1-10; S12), no statistical difference could be determined. No statistical difference could be determined for the fourth part of the survey on AI teaching (S13-20). Regarding the teaching of AI ethics, statistical differences could be determined for 2 statements (S21, \( P=.004 \); S22, \( P=.03 \)). For the statement that the teaching of AI ethics should be part of medical education, the \( z \) score indicated a higher level of agreement in the group that had received prior ethics education. Similarly, a higher level of disagreement was indicated by the group with prior ethics education for the statement that the teaching of AI ethics in medical education is adequate (\( z \) score: -3.011). There was no statistically significant difference in the AI ethics teaching content between the groups. A detailed statistical analysis can be found in Tables S9-S12 in Multimedia Appendix 1.

**Discussion**

This discussion aims to comprehensively analyze the findings regarding medical students’ perceptions of AI in medicine and the role of AI and AI ethics in their medical education, depending on their use of AI-based (chat) applications such as ChatGPT.

**The Use of AI-Based (Chat) Application Among the Surveyed Medical Students**

The discrepancy between students’ personal AI experiences and formal medical education highlights the gap in integrating AI into curricula, reflecting the need for educational progress in line with technological advancement. A considerable 38.8% of the respondents reported prior use of AI-based (chat) applications, such as ChatGPT, Bard, Bing Chat, or Jasper Chat, which was slightly below the percentage received from pretesting and used for sample size calculation (5/11, 45%).

The results concerning the reported use of AI-based (chat) applications must be evaluated in the context of the timing of the data collection. ChatGPT, for instance, became freely available to the public on November 30, 2022, making it accessible for only approximately 8 months at the time of data collection [27]. In addition, Bing Chat was not broadly accessible until May 2023, further constraining its availability before the survey [28]. It is noteworthy that academic literature concerning the use of AI-based (chat) applications such as ChatGPT among medical students is still limited. A study conducted with health students found that only 11.3% (55/458) of respondents reported using the ChatGPT, a rate considerably lower than the findings of this study [29].

A more detailed evaluation of the percentage of medical students using AI-based (chat) applications is necessary given that many might use AI unknowingly. This is not restricted to clinical AI tools, such as clinical decision support systems but extends to search engines and other tools. For example, the search engine Bing offers AI-driven content with search results, irrespective of whether the Bing chat is specifically used. Moreover, a study conducted with students from various specialties in Germany revealed that 12.3% (779/6311) of its participants used “DeepL” (DeepL, SE), an AI-based translation tool, in which the use of AI might not be immediately evident [30]. Therefore, when considering other AI tools and applications, the actual percentage of medical students using them may be significantly higher than the 38.8% reported in this study. Recognizing this potential underestimation of AI use highlights the importance of expanding AI literacy and awareness in medical education to ensure that future health care professionals are adequately prepared for the integration of AI in medicine. This reinforces the need for proactive measures in curriculum design to include not only the direct use of AI tools but also an understanding of their indirect implications in various medical and research contexts.

**AI Education**

Despite the significant engagement of students with AI-based applications, such as ChatGPT, only a small fraction (26/487, 5.3%) reported formal AI education within their medical curriculum. This discrepancy highlights the critical gap between experiential learning and structured academic guidance regarding AI. Interestingly, AI education outside the formal curriculum was more prevalent (51/487, 10.5%), which could imply a proactive approach to learning about AI technologies. Furthermore, this could be attributed to the availability of AI-based applications, such as ChatGPT, and increasing opportunities for education on AI in the medical context, as well as AI-based (chat) applications that are knowledgeable in the field of medicine [7,31-33]. Among the users of AI applications, 73% applied these tools in medical contexts, primarily for querying medical knowledge. This use pattern presents both opportunities for accessible knowledge and risks associated with reliance on uncertified AI sources and a lack of certification as medical devices. The lack of education in the field of AI as part of medical education has been highlighted not only in German-speaking countries [34] but also internationally [21,22].
The results imply a substantial dichotomy between the lack of formal education and optimism toward AI, as the use of AI in medicine was positively perceived (71.1% of respondents), despite the absence of formal education (94.7% of respondents). Given the lack of education, this warrants caution as there might be an overly optimistic view of its potential benefits, overlooking potentially significant limitations and ethical implications [35]. The need for the integration of AI into medical curricula is not only supported by existing studies highlighting low AI literacy among medical students [34,36] but also by the results of this study, with 88% of all medical students perceiving that their current AI education within their medical education is insufficient. This dissatisfaction underscores the need for medical curricula to evolve in tandem with technological advancements. However, it is crucial to ensure that these curricular changes are developed thoughtfully and comprehensively to avoid superficial or overly optimistic portrayals of AI’s role in medicine [34]. The findings of this study, indicating a significant gap in AI education within medical curricula, align with the initial insights gathered regarding students’ use of AI applications. Furthermore, the results align with the objective of understanding how medical students from Germany, Austria, and Switzerland perceive the application of AI in medical practice and its integration into medical education. This disparity between the practical use of AI applications and lack of AI educational opportunities in the curriculum underlines the emerging need for educational reform.

**AI Ethics Education**

The perceived insufficiency of the current medical education extends to AI ethics. Remarkably, 95.3% of participants acknowledged the new ethical challenges posed by AI in medicine, which resonates with preexisting research [15]. Notably, those who used AI-based (chat) applications, such as ChatGPT, agreed more strongly with this view, suggesting that practical use enhances awareness of these ethical issues. In addition, 74.9% (385/487) of respondents recognized the necessity of integrating AI ethics into medical curricula, aligning with recent academic discourse [37-39]. However, only a small percentage (4.3%) reported formal AI ethics education, highlighting a significant deficit in the current curriculum. Medical students perceived all 8 proposed ethical AI topics as highly relevant, which were recommended as potential teaching content for AI ethics in the current literature [37-39]. Statistical differences were observed for “informed consent” among those with prior AI application use. This indicates that engagement with AI technology may deepen understanding of its ethical dimensions, reinforcing the need for comprehensive AI ethics instruction in medical education. The clear demand for AI ethics education reflects a broader educational need, where medical students should not just be prepared for the technicalities of AI but also for the nuanced ethical considerations introduced by the technology.

Although this study underscores the need for both AI and AI ethics education in medical curricula, it is also important to critically assess the current absence of AI-centric content. Rapid technological advancements in AI with the recent public availability of AI tools, such as ChatGPT, may contribute to the current lack of associated teaching content. Given the complex regulatory requirements required to use AI-based technologies in clinical practice, the use of AI in medicine is currently not widespread [40]. In addition, the requirement for time-consuming and complex reaccreditation processes for curricular development and revision may further delay the introduction of AI-related teaching content [41]. Moreover, the lack of widespread use of AI-based applications in medicine and clinical practice likely contributes to the current lack of adequate teaching content on AI and ethics. The overwhelming perception of AI’s potential and its ethical implications it brings forth, as evidenced by this study, underscores the need for educational institutions to respond proactively. Balancing the speed of technological advancements in the field of AI with thoughtful and comprehensive curricular integration is likely to be a crucial challenge in medical education in the coming years.

**Additional Analysis of the Collected Data**

In the additional data analysis, the subsequent examination revealed that perceptions of AI and AI ethics among medical students were not significantly influenced by their country of study. This uniformity across Germany, Austria, and Switzerland suggests consistency in deficiencies in AI and AI ethics education regardless of regional curricular variations. As the findings could be attributed to the limited number of medical students indicating prior education in AI (26/487, 5.3%) and AI ethics (21/487, 4.3%), additional research is warranted. Despite their different educational systems, the observed uniformity in AI and AI ethics education across the 3 countries implies a broader challenge for medical education. The consistency of educational deficiencies, irrespective of regional curricular variations, indicates the widespread need to reform AI teaching in medical curricula. This aligns with the overarching findings of our study, which suggest a universal gap in AI competencies among medical students.

Further analysis of the study stage revealed that students in advanced stages, such as CPS, showed increased awareness of the potential impact of AI on their specialization choices, implying a growing realization of AI’s role as they progress in their studies. However, the lack of significant differences in most other AI-related statements could also imply a generalized consensus or a lack of adequate exposure and understanding across all study stages. As an advancement in the study stages could be linked to statistically significant results on statements regarding the need to teach AI ethics, this could be attributable to prior ethics education, which is usually taught during the PCSs.

The impact of ethics education on perceptions of AI’s role in medicine is particularly notable. Students with such an education showed increased awareness of the ethical challenges posed by AI, especially regarding its potential negative impact on medical staff autonomy (S11). This could underscore the importance of ethics education in understanding the potentially wide-reaching challenges of AI in medicine for ethically important subjects such as autonomy; however, no statistically significant difference for the preceding statement on autonomy “the use of AI in medicine will negatively affect patient autonomy” (S10) could be observed. This could imply that prior ethics education,
including teaching autonomy in a medical context, might lead to a more nuanced understanding of the subject and potential implications of AI. The results of the analysis reinforce the need for ongoing ethics education, not just as a separate entity, but also interwoven with AI-related topics, to enhance students’ comprehensive understanding of the ethical implications of AI in medicine. The significant influence of prior ethics education on shaping students’ perceptions of the role of AI in medicine emphasizes the interaction between ethical training and technological awareness. The nuanced understanding of the ethical implications of AI among students who have received ethics education underscores the importance of such training in developing critical thinking about the impact of AI in healthcare. Integrating ethics education with AI teaching content could foster a more holistic approach, preparing students not only for the technological aspects of AI but also for its ethical and societal implications [37].

Limitations
Despite the strengths of this study, some limitations must be acknowledged. First, our web-based survey could introduce selection bias, as tech-savvy students may be more likely to participate. Second, the survey measured students’ perceptions rather than their actual competencies in AI and ethics. In addition, although estimated, the response rate was suboptimal, which may limit the generalizability of our findings. Geographically, our sample was limited to German-speaking countries, making the translation of these results to other countries with different health care systems and medical educational frameworks difficult. Cultural attitudes toward AI could also vary, possibly influencing students’ perceptions of and engagement with AI. Our study is essentially a snapshot of a rapidly evolving field; hence, our findings may not reflect attitudes and competencies, as they evolve with advancements in AI technology. In our analysis, we observed statistically significant differences based on prior ethics education and study stage. However, although the additional analysis of the data did not show a direct overlap with significant findings between the main and supplementary evaluations, additional tests are needed to determine whether these factors acted as confounders in our main data analysis. Although this study considered specific potential confounders, it is worth noting that other confounding variables may exist and were not analyzed in this study. Finally, owing to the self-reported nature of the data, the responses might be subject to recall bias, misunderstanding of questions, or social desirability bias. Although our findings provide valuable insights into the state of AI education in German-speaking medical schools, broader multinational studies would offer a more comprehensive understanding.

Conclusions
This study provides a valuable understanding of the perceptions and experiences of medical students in Germany, Austria, and Switzerland regarding the application of AI in medicine, and its role in medical education. Our findings clearly indicate a discrepancy between students’ interactions with AI-based chat applications such as ChatGPT and the representation of AI in their formal education. Despite a significant number of students interacting with AI technology, notably AI-based chat applications, only a fraction have received any formal AI education, revealing a substantial gap in the current medical curricula. This highlights the necessity of the evolution of medical curricula to incorporate AI and AI ethics education, ensuring that future medical professionals are adequately equipped to navigate the challenges and opportunities presented by AI in medicine.

Furthermore, our findings indicate that practical engagement with AI technology can contribute to an increased awareness of ethical implications, reinforcing the importance of including hands-on AI experiences in medical education. It is evident that the rapid advancement and application of AI in medicine demand parallel evolution in medical education. Thoughtful and comprehensive curricular changes are required to provide a balanced understanding of the potential benefits, limitations, and ethical implications of AI. The integration of AI and AI ethics into medical education is an urgent necessity, not only to enhance students’ AI literacy but also to ensure the responsible and effective use of AI in future medical practice demands.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Comprehensive statistical analysis and evaluation of confounding factors regarding medical students’ perceptions of artificial intelligence’s role in medicine and medical education.

References


Abbreviations

AI: artificial intelligence
CPS: clinical practical stage
CS: clinical stage
PCS: preclinical stage

©Lukas Weidener, Michael Fischer. Originally published in JMIJ Medical Education (https://mededu.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,
provided the original work, first published in JMI... proper citation. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study

Amanda L Zaleski1,2, MSc, PhD; Rachel Berkowsky3, MSc; Kelly Jean Thomas Craig1, PhD; Linda S Pescatello3, MSc, PhD

1Clinical Evidence Development, Aetna Medical Affairs, CVS Health Corporation, Hartford, CT, United States
2Department of Preventive Cardiology, Hartford Hospital, Hartford, CT, United States
3Department of Kinesiology, University of Connecticut, Storrs, CT, United States

Corresponding Author:
Amanda L Zaleski, MSc, PhD
Clinical Evidence Development
Aetna Medical Affairs
CVS Health Corporation
151 Farmington Avenue
Hartford, CT, 06156
United States
Phone: 1 8605385003
Email: zaleskia@aetna.com

Abstract

Background: Regular physical activity is critical for health and disease prevention. Yet, health care providers and patients face barriers to implement evidence-based lifestyle recommendations. The potential to augment care with the increased availability of artificial intelligence (AI) technologies is limitless; however, the suitability of AI-generated exercise recommendations has yet to be explored.

Objective: The purpose of this study was to assess the comprehensiveness, accuracy, and readability of individualized exercise recommendations generated by a novel AI chatbot.

Methods: A coding scheme was developed to score AI-generated exercise recommendations across ten categories informed by gold-standard exercise recommendations, including (1) health condition–specific benefits of exercise, (2) exercise preparticipation health screening, (3) frequency, (4) intensity, (5) time, (6) type, (7) volume, (8) progression, (9) special considerations, and (10) references to the primary literature. The AI chatbot was prompted to provide individualized exercise recommendations for 26 clinical populations using an open-source application programming interface. Two independent reviewers coded AI-generated content for each category and calculated comprehensiveness (%) and factual accuracy (%) on a scale of 0%-100%. Readability was assessed using the Flesch-Kincaid formula. Qualitative analysis identified and categorized themes from AI-generated output.

Results: AI-generated exercise recommendations were 41.2% (107/260) comprehensive and 90.7% (146/161) accurate, with the majority (8/15, 53%) of inaccuracy related to the need for exercise preparticipation medical clearance. Average readability level of AI-generated exercise recommendations was at the college level (mean 13.7, SD 1.7), with an average Flesch reading ease score of 31.1 (SD 7.7). Several recurring themes and observations of AI-generated output included concern for liability and safety, preference for aerobic exercise, and potential bias and direct discrimination against certain age-based populations and individuals with disabilities.

Conclusions: There were notable gaps in the comprehensiveness, accuracy, and readability of AI-generated exercise recommendations. Exercise and health care professionals should be aware of these limitations when using and endorsing AI-based technologies as a tool to support lifestyle change involving exercise.

(JMIR Med Educ 2024;10:e51308) doi:10.2196/51308

KEYWORDS
exercise prescription; health literacy; large language model; patient education; artificial intelligence; AI; chatbot

https://mededu.jmir.org/2024/1/e51308
Introduction

Regular physical activity is an essential component of a healthy lifestyle with numerous benefits that are widely recognized and indisputable [1,2]. To support overall health, the American College of Sports Medicine (ACSM) and the Department of Health and Human Services recommend healthy adults engage in regular physical activity, including moderate-intensity aerobic exercise for at least 150 minutes per week, vigorous-intensity aerobic exercise for at least 75 minutes per week, or a combination of both, as well as muscle-strengthening activities at least twice per week [1,2]. In addition, evidence-based practice calls for exercise as first-line therapy to prevent, treat, and control multiple chronic conditions and diseases such as hypertension, hypercholesterolemia, and diabetes mellitus [3-7]. As such, the ACSM endorses individualized, evidence-based, exercise recommendations (termed exercise prescription [ExRx]) for more than 25 clinical populations [1]. These ExRxs are tailored to favorably augment health-related outcomes of interest for each respective clinical population while addressing additional factors such as clinical contraindications, common medications, and special considerations [1,8]. Despite well-established guidelines, health care providers often struggle to provide sufficient counseling and follow-up on lifestyle recommendations, including exercise, due to various barriers such as time constraints, limited resources, lack of awareness or training, and lack of reimbursement incentives [9-11]. Patients also rely heavily on web-based sources for health-related information [12-14], which often includes misinformation that can negatively impact health outcomes and undermine provider-led efforts to support behavior change [15,16].

Artificial intelligence (AI) has recently emerged as a promising tool to augment health and health care and address these challenges [17]. AI-based technology including machine learning, neural networking, deep learning, and natural language processing enables computers to interact with a corpus of text data to generate human language [18,19]. Large language models (LLMs), such as the generative pretrained transformer (GPT), have the ability to generate human-like language on their own, making them a powerful tool for interacting with users as if they are communicating with another human [18,19]. The surge in popularity of LLMs can largely be attributed to the third iteration of OpenAI’s GPT series, ChatGPT [20]. ChatGPT has been recognized as the fastest-growing consumer application in history [20] and is widely regarded as disruptive technology due to its strong potential to enable a wide range of clinical applications as both a provider- and patient-facing tool [21] by generating language that is contextually appropriate, natural sounding, and coherent. Indeed, ChatGPT has demonstrated remarkable capabilities including diagnosis support, streamlining clinical workflows, reducing documentation burden, improving patient education understandability and experience [22-25], and, most recently, passing the United States Medical Licensing Examination [26].

Transformative applications of ChatGPT continue to evolve, but evaluation of its output and suitability in clinical context remains to be explored, in addition to identifying barriers to access and outcomes related to its use. The application of digital technology to support a health behavior change using knowledge-shaping techniques, which is complex and riddled with contextual and individualized components, is challenging [27]. Challenges include ensuring the suitability and usability of the technology confers appropriate educational requisites to understand and apply knowledge in the form of its recommendations. These educational considerations include readability, which can influence the use of AI-generated education for health behavior change [28]. Further, as an extension of readability, low health literacy can limit a patient’s ability to understand and use health information effectively, which can reduce the effectiveness of AI-generated educational resources [29,30].

The evaluation of ChatGPT’s suitability to provide interactive, personalized, and evidence-based exercise recommendations to support behavior change to improve health has not been conducted to date. As such, the primary aim of this study is to assess the suitability of exercise recommendations generated by ChatGPT, a new AI chatbot, as an adjuvant educational tool for health care providers and patients. Primary outcomes of interest include comprehensiveness, accuracy, and readability of the recommendations generated by ChatGPT, with the goal of determining its potential to deliver personalized exercise recommendations at scale. A secondary aim of this study was to conduct a qualitative analysis to identify potential patterns, consistencies, and gaps in AI-generated exercise recommendations. As this technology is still nascent, the study was exploratory in nature, without an a priori hypothesis.

Methods

High-Level Overview

This study was conducted in March 2023 using the free research preview of a novel AI chatbot (ChatGPT February 13 version) [31]. Figure 1 provides a conceptual overview of the study. Briefly, open-text queries seeking individualized exercise advice were posed to the chatbot interface for all populations (N=26) for which there exist established evidence-based exercise recommendations by the ACSM [1]. Mixed methods were applied to characterize individual and average exercise recommendation content depth, accuracy, and readability. The results were synthesized to highlight potential strengths, weaknesses, opportunities, and risks for researchers, clinicians, and patients likely to interact with the ChatGPT platform for this use case.
Ethical Considerations

This study was deemed to be exempt by the University of Connecticut Institutional Review Board (E23-0378) as this study solely involved the evaluation of AI-generated output and did not involve interaction or intervention with human subjects.

Selection of the Gold-Standard Reference Source

The ACSM is widely regarded as a leading authority in the field of exercise science and sports medicine, and the organization’s guidelines and recommendations are considered the gold standard for health and fitness professionals in the United States and the world [1,8,32]. The ACSM’s Guidelines for Exercise Testing and Prescription (GETP) serves as its flagship resource manual, continuously updated every 4-5 years since 1975. The most recent edition integrates the latest guidelines from ACSM position stands and other relevant professional organizations’ scientific statements, including the 2018 Physical Activity Guidelines for Americans [1]. This latest edition of GETP represents the most current and primary resource for evidence-based exercise recommendations [1]. Given ACSM’s authoritative status and the comprehensiveness of its guidelines, GETP was selected as the ground truth benchmark source to guide the study design and systematically evaluate the suitability of AI-generated exercise recommendations.

AI-Generated Exercise Recommendations

Following this prompt specificity and structure, all clinical populations within the ACSM GETP were evaluated once in a separate prompt (N=26), including healthy adults, children and adolescents, older adults, persons who are pregnant, and individuals with cardiovascular disease (CVD), heart failure, heart transplant, peripheral artery disease, cerebrovascular accident, asthma, chronic obstructive pulmonary disease, diabetes mellitus, dyslipidemia, hypertension, overweight and obesity, arthritis, cancer, fibromyalgia, HIV, kidney disease, multiple sclerosis, osteoporosis, spinal cord injury, Alzheimer disease, intellectual disability, and Parkinson disease.

Conceptual Content Analysis

A list of conceptual categories was generated, refined, and organized into a coding scheme for predefined categories that pertain to the fundamental aspects of an ExRx. These categories relate to an individualized physical activity program based on the FITT principle, which stands for the frequency (how often?).
intensity (how hard?), time (how long?), and type (what kind?) of exercise [1,35]. The final coding scheme included ten categories: (1) health condition–specific benefits of exercise, (2) exercise preparticipation health screening, (3) frequency, (4) intensity, (5) time, (6) type, (7) volume, (8) progression, (9) special considerations, and (10) references (ie, citations to primary literature or sources that supported the AI-generated content provided).

AI-generated exercise recommendations were then coded and recorded in Microsoft Excel (version 2208; Microsoft Corp) following a 2-stage coding process by 2 independent coders with advanced degrees in kinesiology (ALZ and RB). In the first stage, AI-generated content was appraised for comprehensiveness. Each exercise recommendation was coded for the presence (1 point) or absence (0 points) of content provided for each of the 10 prespecified categories such that each exercise recommendation had a possible range of 0-10 points. Comprehensiveness was determined by dividing the total number of points (ie, actual) by the total number possible (ie, expected or 10 points) and multiplying by 100. The resulting score was expressed as a percent, with 100% indicating the highest possible score and fully comprehensive. This formula was applied to all 26 exercise recommendations and averaged to characterize ChatGPT’s overall ability to deliver exercise recommendations regarding their comprehensiveness.

In the second stage, all categories with reported content (ie, fully and partially comprehensive content) were appraised for accuracy. Accuracy was defined as concordance with the ACSM GETP as the ground truth source [1]. In one instance, content deviated from the ACSM GETP (ie, condition-specific benefits of exercise for individuals with HIV), and accuracy was defined as the degree to which the content was consistent with other widely established facts or clinical literature. Responses were coded by the same independent reviewers (ALZ and RB) and recorded as binary variables: “concordant” or “discordant” following the same process used to determine comprehensiveness. Potential discrepancies in coding were resolved through discussion with a third party and senior expert in the field (LSP). The accuracy score was determined by dividing the number of concordant category counts by the number of categories present (ie, “actual” counts; previously determined when calculating comprehensiveness during the first stage) and multiplying by 100. The resulting score was expressed as a percent, with 100% indicating the highest possible accuracy score or fully concordant.

Readability Metrics
The Flesch-Kincaid formula was used to determine readability, a commonly used tool that evaluates the complexity of text-based educational material. This tool was selected due to its objectivity, as scores are computationally derived rather than paper-and-pencil tools that rely on hand calculations and subjectivity, which introduce risk for human error [36]. The formula is based on the average number of syllables per word and the average number of words per sentence with the resulting score estimating the minimum grade level required to understand the text. For example, a score of 8.0 means that the text can be understood by an average eighth-grader in the United States. Flesch reading ease scores range from 0 to 100, with higher scores indicating easier-to-read text. For example, scores <50 are considered difficult to read, while scores >80 are considered easy to read [36]. To assess readability metrics and word count, a single researcher (RB) used the built-in readability statistics functionality of Microsoft Word (version 2208). The mean (SD) word count and readability metrics (ie, Flesch reading ease and grade level) were calculated using Microsoft Excel (version 2208).

Qualitative Analysis
Qualitative analysis with a thematic mapping approach was used to identify novel patterns, trends, and insights across the AI-generated text output. Thematic mapping, a qualitative research method, involves the identification, analysis, and visualization of recurring themes or topics within a data set. This approach is instrumental in highlighting consistencies or gaps in data, facilitating the generation of insights, and formulating hypotheses for further investigation [37].

Statistical Analyses
Descriptive statistics characterized the distribution of all outcome variables of interest, including comprehensiveness, accuracy, and readability metrics. Interrater reliability was assessed using Cohen \( \kappa \) coefficient [(observed agreement–expected agreement)/(1–expected agreement)]. Qualitative analysis was conducted using a systematic multistep approach. All AI-generated exercise recommendations, comprising the text output, were collected and organized to form the data set for qualitative examination. The analysis was carried out by a single researcher (ALZ) who immersed themselves in the content and initiated the coding process by identifying initial themes or patterns within the recommendations. Subsequently, codes were meticulously refined and organized into broader themes, ensuring consistency and accuracy throughout the process. These identified themes were then visually mapped to represent patterns within the data set. Insights generated from the analysis were discussed collaboratively as a team, facilitating comprehensive understanding and quantification, whenever applicable.

Results

Interrater Reliability
Interrater reliability was assessed for the 2 independent raters who coded a sample of 26 AI-generated exercise recommendations using a set of 10 categories. Cohen \( \kappa \) coefficient was calculated to be 1.0, indicating perfect agreement between coders.

Comprehensiveness of AI-Generated Exercise Recommendations
Table 1 details the presence of educational content across the predefined categories of interest abstracted from AI-generated exercise recommendations for 26 populations. Overall, AI-generated exercise recommendations were 41.2% (107/260) comprehensive when compared against a predefined set of content categories that comprise a gold-standard ExRx [1]. There were no populations or categories that were fully...
Comprehensiveness ranged from 0% to 92% with notable gaps in content surrounding the critical components of ExRx: frequency (n=2, 8%), intensity (n=2, 8%), time (n=1, 4%), and volume (n=0, 0%). Partial information was provided across these same categories (ranging from 31% to 58%) with almost all gaps surrounding the provision of FITT for resistance training or flexibility modalities. In addition, only 8% (n=2) of recommendations provided a reference source, both of which (accurately) cited the American Heart Association.

### Table 1. Comprehensiveness of artificial intelligence–generated exercise recommendations by content category (N=26).

<table>
<thead>
<tr>
<th>Content</th>
<th>Fully provided, n (%)</th>
<th>Partiala, n (%)</th>
<th>Not provided, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition-specific benefits</td>
<td>24 (92)</td>
<td>0 (0)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>Preparticipation screening</td>
<td>24 (92)</td>
<td>0 (0)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>Frequency</td>
<td>2 (8)</td>
<td>9 (35)</td>
<td>15 (58)</td>
</tr>
<tr>
<td>Intensity</td>
<td>2 (8)</td>
<td>15 (58)</td>
<td>9 (35)</td>
</tr>
<tr>
<td>Time</td>
<td>1 (4)</td>
<td>10 (38)</td>
<td>15 (58)</td>
</tr>
<tr>
<td>Type</td>
<td>14 (54)</td>
<td>12 (46)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Volume</td>
<td>0 (0)</td>
<td>8 (31)</td>
<td>18 (69)</td>
</tr>
<tr>
<td>Progression</td>
<td>15 (58)</td>
<td>0 (0)</td>
<td>11 (42)</td>
</tr>
<tr>
<td>Special considerations</td>
<td>23 (88)</td>
<td>0 (0)</td>
<td>3 (12)</td>
</tr>
<tr>
<td>References</td>
<td>2 (8)</td>
<td>0 (0)</td>
<td>24 (92)</td>
</tr>
</tbody>
</table>

aPartial indicates some, but not all, possible content was provided.

**Accuracy of AI-Generated Exercise Recommendations**

Of the total available content provided to the end user, AI-generated exercise recommendations were 90.7% (146/161) accurate when compared to a gold-standard reference source (ie, ACSM GETP [1]). Among the 9.3% (15/161) of inaccurate recommendations (Table 2), there were 15 counts of discordance with most misinformation counts (n=8, 53%) surrounding the need for preparticipation medical clearance prior to engaging in exercise. The second highest category of discordance was within education related to frequency (n=2, 13%) with “overprescribing” aerobic exercise for Alzheimer disease and fibromyalgia by 2 and 5 days per week, respectively. There was 1 count each of discordance across 5 content categories (ie, condition-specific benefits, intensity, time, type, and progression) and 0 counts of discordance across the remaining content categories, including volume, special considerations, and references.

When comparing populations with discordance, hypertension (n=3) had the greatest number of misinformation counts followed by individuals with fibromyalgia (n=2), healthy adults (n=1), older adults (n=1), and cancer (n=1) and individuals with Alzheimer disease (n=1), arthritis (n=1), dyslipidemia (n=1), HIV (n=1), multiple sclerosis (n=1), osteoporosis (n=1), and overweight and obesity (n=1).
Table 2. Summary of inaccurate content among AI-generated exercise recommendations for all reported content categories (N=161).

<table>
<thead>
<tr>
<th>Content category (count/report-ed) and population</th>
<th>AI-generated content</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition-specific benefits (1/24)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV</td>
<td>“Regular exercise can be beneficial for people living with HIV. It can help improve cardiovascular health, reduce stress and anxiety, boost immune function, and enhance overall well-being.”</td>
<td>Research regarding the impact of exercise to improve immune function is inconclusive [38].</td>
</tr>
<tr>
<td>Preparticipation screening (8/24)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy adult</td>
<td>“Remember to always listen to your body and consult with a healthcare professional before starting a new exercise routine.”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Older adult</td>
<td>“However, it’s important for older adults to choose exercises that are appropriate for their age and fitness level, and to check with their doctor before starting a new exercise program.”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>“It’s important to consult with a healthcare professional before starting a new exercise program...”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Hypertension</td>
<td>“Consult with your doctor: Before starting any exercise program, it is important to consult with your doctor to make sure it is safe for you to exercise and to get guidance on the most appropriate type and intensity of exercise.”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Arthritis</td>
<td>“Be sure to talk to your doctor or a physical therapist before starting any new exercise program...”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Cancer survivors</td>
<td>“It’s important to work with a healthcare provider or a qualified exercise professional to develop a safe and effective exercise program that meets your individual needs and abilities.”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>“It is always recommended to consult with a healthcare professional before starting any exercise program.”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>“It’s important to talk to your doctor or a qualified exercise professional before starting any new exercise program, especially if you have osteoporosis or other medical conditions.”</td>
<td>ACSM preparticipation screening guidelines state that “physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines” [1].</td>
</tr>
<tr>
<td>Frequency (2/11)</td>
<td>“Aim for at least 30 minutes of aerobic exercise most days of the week.”</td>
<td>ACSM recommends an initial frequency of 1-2 days per week, gradually progressing to 2-3 days per week [1].</td>
</tr>
<tr>
<td>Content category (count/report-ed) and population</td>
<td>AI-generated content</td>
<td>Ground truth</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td>----------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Alzheimer disease</td>
<td>“Engage in moderate aerobic exercise such as brisk walking, cycling, or swimming for at least 30 minutes a day, five days a week.”</td>
<td>ACSM recommends a frequency of 3 days per week [1].</td>
</tr>
<tr>
<td>Intensity (1/17)</td>
<td>“Avoid high-intensity exercises: Avoid high-intensity exercises that can cause sudden increases in blood pressure, such as sprinting or heavy lifting.”</td>
<td>ACSM does not contraindicate vigorous-intensity aerobic exercise or heavy lifting assuming adequate progression, absence of underlying disease, and proper breathing technique (ie, avoidance of Valsalva maneuver) [1].</td>
</tr>
<tr>
<td>Hypertension</td>
<td>“Start with 1-2 sets of 10-15 repetitions for each exercise and gradually increase the resistance as tolerated.”</td>
<td>ACSM recommends gradual progression of 4-5 to 8-12 repetitions and increasing from 1 to 2-4 sets per muscle group [1].</td>
</tr>
<tr>
<td>Time (1/11)</td>
<td>“Aim for at least 30 minutes of moderate-intensity aerobic exercise most days of the week.”</td>
<td>New ACSM guidelines reinforce that emphasis is no longer placed on aerobic exercise alone. Aerobic or resistance exercise alone or aerobic and resistance exercise combined (ie, concurrent exercise) is recommended on most, preferably all, days of the week to total 90 to 150 minutes per week or more of multimodal, moderate-intensity exercise [39].</td>
</tr>
<tr>
<td>Fibromyalgia</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Type (1/26)</td>
<td>“If you’re new to exercise, start with low-intensity activities such as walking or swimming, and gradually increase your intensity and duration.”</td>
<td>ACSM recommends initial intensity should be moderate, progressing to vigorous for greater health benefits [1].</td>
</tr>
<tr>
<td>Hypertension</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Volume (0/8)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Progression (1/15)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Overweight and obesity</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Special considerations (0/23)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>References (0/2)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

aAI: artificial intelligence.
bACSM: American College of Sports Medicine.
cN/A: not applicable.

**Readability Metrics**

Average and individual readability metrics and word count for AI-generated exercise recommendations are provided in Table 3. On average, AI-generated output was 259.3 (SD 49.1) words (range 171-354) and considered “difficult to read” with an average Flesch reading ease of 31.1 (SD 7.7; range 14.5-47.3) and written at a college-level (mean 13.7, SD 1.7; range 10.1-18.0).
Table 3. Readability metrics for artificial intelligence–generated exercise recommendations by population.

<table>
<thead>
<tr>
<th>Population</th>
<th>Word count</th>
<th>Flesch reading ease</th>
<th>Grade level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy adults</td>
<td>187</td>
<td>14.5</td>
<td>15.2</td>
</tr>
<tr>
<td>Children and adolescents</td>
<td>253</td>
<td>29.8</td>
<td>14.1</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>267</td>
<td>34.7</td>
<td>13.5</td>
</tr>
<tr>
<td>Older adults</td>
<td>276</td>
<td>37.0</td>
<td>12.2</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>271</td>
<td>33.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Heart failure</td>
<td>235</td>
<td>23.0</td>
<td>16.2</td>
</tr>
<tr>
<td>Heart transplant</td>
<td>278</td>
<td>24.9</td>
<td>14.4</td>
</tr>
<tr>
<td>Peripheral artery disease</td>
<td>322</td>
<td>32.4</td>
<td>13.4</td>
</tr>
<tr>
<td>Cerebrovascular accident</td>
<td>346</td>
<td>22.0</td>
<td>15.1</td>
</tr>
<tr>
<td>Asthma</td>
<td>317</td>
<td>41.1</td>
<td>12.0</td>
</tr>
<tr>
<td>COPD(^a)</td>
<td>247</td>
<td>47.3</td>
<td>10.1</td>
</tr>
<tr>
<td>Diabetes</td>
<td>201</td>
<td>36.7</td>
<td>11.8</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>291</td>
<td>19.6</td>
<td>15.9</td>
</tr>
<tr>
<td>Hypertension</td>
<td>247</td>
<td>34.5</td>
<td>13.3</td>
</tr>
<tr>
<td>Overweight and obesity</td>
<td>200</td>
<td>34.7</td>
<td>13.2</td>
</tr>
<tr>
<td>Arthritis</td>
<td>236</td>
<td>38.4</td>
<td>13.0</td>
</tr>
<tr>
<td>Cancer</td>
<td>319</td>
<td>24.8</td>
<td>14.9</td>
</tr>
<tr>
<td>Fibromyalgia</td>
<td>303</td>
<td>40.0</td>
<td>12.2</td>
</tr>
<tr>
<td>HIV</td>
<td>232</td>
<td>30.0</td>
<td>13.9</td>
</tr>
<tr>
<td>Kidney disease</td>
<td>354</td>
<td>31.1</td>
<td>15.3</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>255</td>
<td>38.4</td>
<td>11.4</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>171</td>
<td>32.7</td>
<td>12.3</td>
</tr>
<tr>
<td>Spinal cord injury</td>
<td>281</td>
<td>25.5</td>
<td>14.1</td>
</tr>
<tr>
<td>Alzheimer disease</td>
<td>191</td>
<td>29.1</td>
<td>14.8</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>241</td>
<td>32.1</td>
<td>13.2</td>
</tr>
<tr>
<td>Parkinson disease</td>
<td>221</td>
<td>19.8</td>
<td>18.0</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>259.3 (49.1)</td>
<td>31.1 (7.7)</td>
<td>13.7 (1.7)</td>
</tr>
</tbody>
</table>

\(^a\)COPD: chronic obstructive pulmonary disease.

Qualitative Analysis

A secondary aim of this study was to identify potential patterns, consistencies, and gaps in AI-generated exercise recommendation text outputs. Major observations derived from qualitative evaluation of AI-generated exercise recommendations can be found in Multimedia Appendix 2. Briefly, several recurring themes emerged among the total sample, including liability and safety, preference for aerobic exercise, and inconsistencies in the terminology used for exercise professionals. Importantly, AI-generated output showed potential bias and discrimination against certain age-based populations and individuals with disabilities. The implications of these findings are discussed in detail below.

Discussion

Principal Findings

This study sought to explore the suitability of AI-generated exercise recommendations using a popular generative AI platform, ChatGPT. Given the recent launch and popularity of ChatGPT and other similar generative AI platforms, the overall goal was to formally appraise the suitability and readability of AI-generated output likely to be seen by patients and inform exercise and health care professionals and other stakeholders on the potential benefits and limitations of using AI to leverage for patient education. The major findings were that AI-generated output (1) presented 41.2% (107/260) of the content provided in a gold-standard exercise recommendation indicating poor comprehensiveness; (2) of the content provided, chat output was 90.7% (146/161) accurate with most discordance related...
to the need for exercise preparticipation health screening; and (3) had college-level readability.

The results of this study are consistent with a recently published research letter that evaluated the appropriateness of CVD prevention recommendations from ChatGPT [40]. Sarraju et al [40] developed 25 questions on fundamental heart disease concepts, posed them to the AI interface, and subjectively graded responses as “appropriate” or “inappropriate.” AI-generated responses were deemed to be 84% appropriate with noted misinformation provided for questions surrounding ideal exercise volume and type for heart and heart disease prevention. This study expands upon these findings by focusing on ExRx, testing additional metrics (ie, comprehensiveness and readability) using an objective, formal coding system based on a ground truth source, and in an expanded list of clinical populations.

Real-World Implications of These Findings

Our findings suggest that while AI-generated exercise recommendations are generally accurate (146/161, 90.7%), they may lack comprehensiveness in certain critical components of ExRx such as target frequency, intensity, time, and type of exercise, which could potentially hinder ease of implementation or their effectiveness. The most common (ie, 8/15, 53%) source of misinformation was the recommendation to seek medical clearance prior to engaging in any exercise. Potential downstream implications are undue patient concern and triggering an unnecessary number of adults for medical evaluation, both posing as potential barriers to exercise adoption [41,42].

The ACSM preparticipation screening guidelines emphasize the public health message that exercise is important for all individuals and that the preparticipation health screening should not be a deterrent to exercise participation [41]. The preparticipation screening algorithm considers current physical activity levels, desired exercise intensity, and the presence of known or underlying CVD, metabolic, and renal disease. Following this algorithm, lesser than 3% of the general population would be referred before beginning vigorous exercise, and approximately 54% would be referred before beginning any exercise [42]. Interestingly, exercise professionals are well-equipped to facilitate preparticipation screening, yet AI-generated output disproportionately emphasized medical clearance by a health care provider or doctor prior to working with an exercise professional. In reference to exercise professionals, ChatGPT used varying and incorrect terminology such as “licensed exercise physiologist” that does not reflect current-state credentialing for exercise professionals working with clinical populations (ie, ACSM Certified Clinical Exercise Physiologist [43]). These findings corroborate with existing challenges in the public health’s understanding of the role of exercise professionals, levels of qualification, and respective scope of practice [44].

As AI-based technologies continue to evolve, striking the right balance between medical precision and risk mitigation remains a crucial consideration [45]. The question of how definitive an AI-based model should be when delivering medical education is multifaceted. On the one hand, the inclination of the AI-based model toward vague or general recommendations can be seen as a responsible stance to mitigate risks. On the other hand, there is merit in AI-based models providing clear, specific, and contextual guidance that reinforces evidence-based recommendations. This approach ensures that end users receive accurate and tailored advice, which is important in the context of medical education. This tension highlights the need for continued dialogue on how AI can enhance health care while ensuring that recommendations align with the highest standards of accuracy and patient safety. These discussions will be instrumental in shaping the future of AI-augmented health care.

AI-Generated Output Least Accurate for Populations With Hypertension

Interestingly, the hypertension exercise recommendations scored the poorest (ie, highest discordance) with 57% (4/7) accuracy and misinformation surrounding the need for medical clearance and the recommended intensity and type of exercise (Table 2). For example, AI-generated output recommended avoiding high-intensity exercise “such as sprinting or heavy lifting”; however, the ACSM does not contraindicate vigorous-intensity exercise considering comorbidities and assuming adequate progression and proper technique [1]. Additionally, AI-generated output recommended a target exercise goal of “30 minutes of moderate-intensity aerobic exercise most days of the week.” Notably, the ACSM guidelines reinforce that emphasis is no longer placed on aerobic exercise alone but rather recommend aerobic and resistance exercise alone or combined (ie, concurrent exercise) on most, preferably all, days of the week to total 90-150 minutes per week or more of multimodal, moderate-intensity exercise [39]. Reasons for this discordance are likely because the ChatGPT model relies on training data preceding 2021 and may not capture real-time research advancements. Nevertheless, these findings are important because hypertension is the most common, costly, and modifiable CVD risk factor with strong evidence-based and guideline-driven recommendations, whereby support of exercise is a critical component of first-line treatment for elevated blood pressure [7,46-48].

Social Determinants of Health Considerations

Not surprisingly, our evaluation of this AI-based technology identified social determinants of health considerations regarding educational obtainment for its users. Average readability of the AI-generated output was found to be very high, at the college level, which poses significant challenges for the majority of patients, as The National Institutes of Health, American Medical Association, and American Heart Association all recommend that patient education materials be written at or below a sixth-grade reading level [49] based on national educational obtainment trends. Poor readability of patient materials can exacerbate disparities in access to care for those with limited health literacy, and those individuals may experience more barriers to understand and apply the information provided [29,30]. These findings highlight the need for ongoing evaluation and refinement of AI-generated educational output to prevent inappropriate recommendations that do not improve disparities in clinical outcomes. AI-based models, such as ChatGPT, and their output are vulnerable to both poor data

https://imedex.jmir.org/2024/11/e51308
quality and noninclusive design. Notably, AI-generated output used different tenses and pronouns depending on the demographic group being addressed, which potentially perpetuates digital discrimination including stereotypes and biases (Multimedia Appendix 2). For instance, most AI-generated exercise recommendations were provided in the second-person tense; however, recommendations for individuals with intellectual disabilities, older adults, and children and adolescents were written in the third-person tense with the AI-based model, assuming these populations were not the primary end users. Additionally, most exercise examples provided by the chatbot were activities favoring ambulating individuals (eg, walking and running) potentially limiting education for, and perpetuating bias against, individuals with disabilities. Generative AI can contribute to bias or discrimination in several ways, beginning with the use of biased data to train AI-based models that learn and perpetuate biases in its output [50]. Additionally, AI-based models may be designed with certain features that result in biased or discriminatory outputs, such as using certain variables that are correlated with gender or race [50]. Put in practice, AI-based models can further extend societal biases and stereotypes by relying on existing patterns and trends in the data that reinforce gender or racial stereotypes [50]. These findings highlight the need for caution in using generative AI for health education and the importance of careful consideration of potential biases and discriminatory language.

To summarize, this study demonstrates that AI-generated exercise recommendations hold some promise in accurately providing exercise information but are not without issues (ie, gaps in critical information, biases, and discrimination) that could lead to potentially harmful consequences. The art of ExRx involves considering individual factors and nuances that may not be fully captured by technology [1]. Factors such as medical history, medications, personal preferences, health and physical literacy, and physical limitations are just a few examples of the complexities involved in creating an individualized exercise plan [1]. It is important to note that AI-generated output often lacks references to primary sources or literature, underscoring the need for health care provider oversight in interpreting and verifying the validity of the information presented. In this study, the reference sources provided were 100% accurate (2 of 2); however, “hallucinations” of fabricated or inaccurate references are quite common and are a growing concern for AI-generated medical content [51].

Limitations

There are limitations to this study. This evaluation was limited to a single generative AI platform, which may not be representative of all LLM programs. Additionally, this study is limited to a specific time period and topic, and the findings may not be generalizable to other topics or time periods. Importantly, this model was evaluated using a single, structured prompt that can potentially lead to overfitting or superficial outputs and compromise generalizability. The lack of exposure to a range of prompts makes it challenging to discern if outcomes truly reflect the model’s capabilities or are specific to the nature of the provided prompt. Given that LLMs can yield varied outcomes based on prompts, this limitation is critical for the interpretation and application of the model’s results across various scenarios. This approach was selected as it most closely recapitulates how a publicly available chatbot would likely be used in a real-world setting by an inexperienced end user (ie, lacking knowledge of prompt methodologies). Indeed, all (N=26) AI-generated exercise recommendations were coherent, contextual, and relevant suggesting that the standardized single prompt was structured to elicit an appropriate response. However, it is likely that additional prompt engineering considerations (ie, specificity, iteration, and roles and goals) will yield incremental capabilities and superior model performance than reported in this study. Future work should consider advanced and diverse prompts to assess the model’s robustness across various scenarios. The results rely on the accuracy of the coders in identifying relevant content and assessing its accuracy. The high level of agreement between raters suggests that the coding scheme was well-defined and easily interpretable; however, there is potential for observer bias due to the raters’ shared mentorship, research training, and educational experiences. It is also worth noting that this study used the Flesch-Kincaid formula to assess readability that has known limitations, such as not accounting for the complexity of ideas and vocabulary and not considering readers’ cultural and linguistic backgrounds [36]. This tool was selected due to its objectivity, standardization, and the fact that scores are computationally derived, which lowers the risk of human error, thus rendering it the most appropriate tool to address this research question [36]. Nevertheless, future research may benefit from examining the Flesch-Kincaid formula in conjunction with other measures to gain a more comprehensive understanding of AI-generated output readability.

Despite the noted limitations, this study possesses several strengths. To the best of our knowledge, this study is the first to report on the quality of AI-generated exercise recommendations for individuals across the life span (ie, children and adolescents, healthy adults, and older adults) and for 23 additional clinical populations. A major strength of this study is the use of a formal grading framework with a double-coding system to objectively assess the comprehensiveness and accuracy of the AI-generated exercise recommendations, which extends the literature and increases the reliability and validity of these findings [40]. Adding to its credibility, this grading system was developed and refined by experts in the field of exercise science, including a former associate editor [35], editor, and contributing author [1] of the ACSM GETP (LSP and ALZ). Multiple measures were used to assess the suitability of AI-generated recommendations and its potential for digital discrimination. Recommendations were evaluated by their comprehensiveness, accuracy, and readability, which provided a thorough summarization of the strengths and weaknesses of AI-generated content. The output was compared to well-established evidence-based guidelines (ie, ACSM GETP) as a gold-standard reference, which strengthens the validity of the results. Finally, the standardization of queries in this study minimized bias and allowed for an objective evaluation of the AI-generated exercise recommendations. These structured prompts were integral to the research design, shaping the language model’s responses and enabling the systematic evaluation of its performance against ACSM GETP as the
ground truth benchmark. This methodological approach ensures that the outcomes presented in this study are grounded in a consistent and rigorously designed interaction process.

Future Directions

Given the recent development of open-source generative AI technologies, this area is ripe for exploration. However, before proceeding with extensive randomized controlled trials, it is crucial to prioritize the safety and ethical considerations associated with AI-generated medical education. As AI technologies have the potential to impact health disparities, it is essential to carefully evaluate their use to ensure inclusivity and appropriate messaging across demographics [27,52-54]. Further research is needed to develop, test, and implement AI technologies that serve individuals safely, effectively, and ethically without perpetuating bias, discrimination, or causing harm. This includes exploring ways to mitigate potential biases and discriminatory outcomes. Outside of the research setting, health care and exercise professionals can play a crucial role in improving AI-based models through prompting and by giving corrective feedback to retrain biases and inaccuracies in AI-generated responses. By enriching ChatGPT with user-specific data including exercise components, literacy level, physical limitations, and other activity considerations, there are opportunities to improve the personalization of recommendations and lessen digital discrimination. Through this stewardship, continuous refinement will likely improve the performance, usability, and appropriateness of the model, translating to superior patient outcomes, which is the goal of provider-enablement and patient-facing tools. As LLMs continue to evolve, it will become increasingly important for researchers to continuously assess improvements with response variations over time. Importantly, future work should explore the incremental value of advanced and diverse prompting considerations. Examples of prompting considerations include the provision of roles and goals (eg, “You are a Clinical Exercise Physiologist and your goal is to design a safe and effective exercise prescription to lower blood pressure”), engaging in multiple or chain prompting and specifically prompting for content commonly missing from output as identified in this study.

To ensure the responsible and safe deployment of AI technologies in health care, conducting thorough implementation studies is a logical next step. These studies should focus on measuring various factors, including acceptability, adoption, appropriateness, costs, feasibility, fidelity, penetration, and sustainability. By thoroughly investigating these implementation aspects, we can ensure that the technology is well-integrated and does not pose any harm to patients or health care systems. Following the completion of the implementation studies, it is important to assess the impact of AI-generated models on service outcomes. This includes evaluating health care quality factors such as safety, timeliness, efficiency, effectiveness, equity, and patient-centeredness [55]. Understanding how AI technologies influence these service outcomes will provide valuable insights into their overall impact on health care delivery. Additionally, measuring patient-centered and end-user outcomes is essential to evaluate the effectiveness of AI technologies in improving patient experiences and outcomes. Randomized controlled trials designed to test ChatGPT as an intervention to augment behavior change and associated health outcomes would be of great public health interest. These trials should prioritize patient-centered outcomes, including satisfaction, usability, experience, and patient activation [56]. By assessing these outcomes, we can determine the effectiveness of AI technologies in empowering patients and fostering meaningful engagement with health care providers.

Conclusions

To conclude, this study found that AI-generated exercise recommendations have moderate comprehensiveness and high accuracy when compared to a gold-standard reference source. However, there are notable gaps in content surrounding critical components of ExRx and potentially biased and discriminatory outputs. Additionally, the readability level of the recommendations may be too high for some patients, and the lack of references in AI-generated content may be a significant limitation for use. Health care providers and patients may wish to remain cautious in relying solely on AI-generated exercise recommendations and should limit their use in combination with clinical expertise and oversight.

Acknowledgments

This study was supported by the University of Connecticut, CVS Health Corporation, and Hartford Hospital.

Authors' Contributions

ALZ contributed to the study conceptualization, project management, study design, data curation and coding, statistical analysis, interpretation of the data, visual presentation of the data, and paper preparation and submission. RB contributed to the study design, data coding, interpretation of the data, and copyediting of the paper. KJTC contributed to the interpretation of the data, business leadership, and copyediting of the paper. LSP contributed to the study design, project oversight, interpretation of the data, and revising and copyediting of the paper. All authors contributed to the writing of the paper, reviewed and approved the final version of the paper, and agreed with the order of presentation of the authors.

Conflicts of Interest

ALZ and KJTC are both employed and hold stock with CVS Health Corporation. This study is an objective evaluation to better understand ChatGPT and its outputs. To the best of our knowledge, CVS Health does not currently use or endorse the use of ChatGPT for lifestyle recommendations. LSP is the sole proprietor and founder of P3-EX, LLC, which could potentially benefit
from the results used in this research. The results of this study do not constitute endorsement by the American College of Sports Medicine.

Multimedia Appendix 1
Output from artificial intelligence–generated exercise recommendations for clinical populations (N=26).

[PDF File (Adobe PDF File), 243 KB - mededu_v10i1e51308_app1.pdf ]

Multimedia Appendix 2
Summary of major themes derived from artificial intelligence–generated exercise recommendations.

[PDF File (Adobe PDF File), 116 KB - mededu_v10i1e51308_app2.pdf ]

References


Abbreviations

ACSM: American College of Sports Medicine
AI: artificial intelligence
CVD: cardiovascular disease
ExRx: exercise prescription
FITT: frequency, intensity, time, and type
GETP: Guidelines for Exercise Testing and Prescription
GPT: Generative Pre-trained Transformer
LLM: Large Language Model

44. Gallo PM. The United States Registry for Exercise Professionals: how it works and ways it can advance the fitness profession. ACSM’s Health Fitness J 2023;27(2):51-53. [doi: 10.1249/hfi.0000000000002084]
© Amanda L Zaleski, Rachel Berkowsky, Kelly Jean Thomas Craig, Linda S Pescatello. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators' Perspectives

Yaser Mohammed Al-Worafi¹,², PhD; Khang Wen Goh³, PhD; Andi Hermansyah⁴, PhD; Ching Siang Tan⁵, PhD; Long Chiau Ming⁶, PhD

¹College of Medical Sciences, Azal University for Human Development, Sana’a, Yemen
²College of Pharmacy, University of Science and Technology of Fujairah, Fujairah, United Arab Emirates
³Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia
⁴Department of Pharmacy Practice, Faculty of Pharmacy, Universitas Airlangga, Surabaya, Indonesia
⁵School of Pharmacy, KPJ Healthcare University, Nilai, Malaysia
⁶School of Medical and Life Sciences, Sunway University, Selangor, Malaysia

Corresponding Author:
Ching Siang Tan, PhD
School of Pharmacy
KPJ Healthcare University
Lot PT 17010 Persiaran Seriemas
Kota Seriemas
Nilai, 71800
Malaysia
Phone: 60 67942692
Email: tcsiang@kpju.edu.my

Abstract

Background: Artificial Intelligence (AI) plays an important role in many fields, including medical education, practice, and research. Many medical educators started using ChatGPT at the end of 2022 for many purposes.

Objective: The aim of this study was to explore the potential uses, benefits, and risks of using ChatGPT in education modules on integrated pharmacotherapy of infectious disease.

Methods: A content analysis was conducted to investigate the applications of ChatGPT in education modules on integrated pharmacotherapy of infectious disease. Questions pertaining to curriculum development, syllabus design, lecture note preparation, and examination construction were posed during data collection. Three experienced professors rated the appropriateness and precision of the answers provided by ChatGPT. The consensus rating was considered. The professors also discussed the prospective applications, benefits, and risks of ChatGPT in this educational setting.

Results: ChatGPT demonstrated the ability to contribute to various aspects of curriculum design, with ratings ranging from 50% to 92% for appropriateness and accuracy. However, there were limitations and risks associated with its use, including incomplete syllabi, the absence of essential learning objectives, and the inability to design valid questionnaires and qualitative studies. It was suggested that educators use ChatGPT as a resource rather than relying primarily on its output. There are recommendations for effectively incorporating ChatGPT into the curriculum of the education modules on integrated pharmacotherapy of infectious disease.

Conclusions: Medical and health sciences educators can use ChatGPT as a guide in many aspects related to the development of the curriculum of the education modules on integrated pharmacotherapy of infectious disease, syllabus design, lecture notes preparation, and examination preparation with caution.

(JMIR Med Educ 2024;10:e47339) doi:10.2196/47339

KEYWORDS

innovation and technology; quality education; sustainable communities; innovation and infrastructure; partnerships for the goals; sustainable education; social justice; ChatGPT; artificial intelligence; feasibility
Introduction

Artificial intelligence (AI) plays an important role nowadays rather than at any time in history in many fields, including medical education, practice, and research [1-6]. AI can be defined as the “science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable” [7], or as “a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behaviour, and with the creation of artefacts that exhibit such behaviour” [8]. One of the recent advances in AI development is the launch of a model called ChatGPT, which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests; ChatGPT is a general large language model (LLM) developed recently by OpenAI. While the previous class of AI models have primarily been deep learning models, which are designed to learn and recognize patterns in data, LLMs are a new type of AI algorithm trained to predict the likelihood of a given sequence of words on the basis of the context of the words that appear before it [9].

Empirical studies have demonstrated the effectiveness of AI-based educational tools in various domains. Recent research published in *JMIR Medical Education* [10] on February 8, 2023, evaluated ChatGPT’s potential as a medical education instrument. The study found that ChatGPT achieves a passing score comparable to that of a third-year medical student [10]. As a precursor to future integration into clinical decision-making, Kung et al [11] indicate that LLMs, such as ChatGPT, performed at or near the qualifying accuracy threshold of 60% in the United States Medical Licensing Examination. Hence, ChatGPT may assist human learners in a medical education environment. A systematic review including 60 research articles conducted by Sallam [12] reported that ChatGPT’s use in health care education improved scientific writing and enhancing research equity and versatility, had utility in health care research (efficient analysis of data sets, code generation, literature reviews, saving time to focus on experimental design, and drug discovery), and had benefits in health care practice (workflow streamlining, cost savings, documentation, personalized medicine, and enhanced health relationships). Many educators, researchers, health care professionals and students started using ChatGPT at the end of 2022 for many purposes, such as preparing lecture notes, assignments, literature reviews, and others. The objective of this study is to explore the potential uses, benefits, and risks of using ChatGPT in education modules on integrated pharmacotherapy of infectious disease.

Methods

Study Design

A content analysis of the potential applications of the ChatGPT model for education modules on integrated pharmacotherapy of infectious disease was performed. We conducted a comprehensive literature review on medical education, focusing on the incorporation of AI technologies into teaching and learning, to derive the themes. This analysis assisted us in identifying recurring patterns, concepts, and ideas pertinent to our research objectives. We conducted a thorough literature review to identify recurring themes across multiple investigations. These themes served as the basis for our discussion and analysis. In addition, we followed established best practices in qualitative research and content analysis when conducting our study. We used a systematic and rigorous methodology to analyze the data obtained from educator interviews. Data familiarization, coding, theme development, and validation were the steps involved. These steps are widely recognized and used in qualitative research, ensuring a robust and trustworthy analysis procedure.

Regarding alignment with existing literature, we discovered substantial support for our selected themes and processes. Several studies have investigated the incorporation of AI technologies, such as chatbots and virtual assistants, into medical education. Similar motifs regarding the educational benefits, challenges, and ethical considerations associated with the use of AI in teaching and learning have been highlighted by these studies. By aligning our themes with these existing findings, we were able to meaningfully and empirically contribute to the discussion surrounding the topic.

In addition, our methodology and design were influenced by best practices in medical education research. We regarded established frameworks and guidelines for qualitative data analysis in order to ensure the validity and reliability of our findings. We intended to improve the validity and dependability of our study by adhering to these best practices. Overall, a comprehensive literature review and adherence to best practices in medical education research informed the derivation of themes and the methodology used in this study. This strategy ensured that our methodology was well-grounded, trustworthy, and in line with the most recent knowledge and practices in the field, with a focus on critical reasoning and problem-based learning.

Data Collection

Overview

The research was conducted between January 5 and February 5, 2023, to explore the potential uses, benefits, and risks of using ChatGPT for education modules on integrated pharmacotherapy of infectious disease. Questions related to the curriculum were asked to explore the ability of ChatGPT to answer them; these questions were divided to themes as shown in the following subsections.

Theme 1

Questions related to the development of the curriculum of the education modules on integrated pharmacotherapy of infectious disease, as suggested by Thomas et al [13], were included in accordance with the following 6 steps: (1) step 1: problem identification and general needs assessment; (2) step 2: targeted needs assessment; (3) step 3: goals and objectives; (4) step 4: educational strategies; (5) step 5: implementation (not included herein); and (6) step 6: evaluation and feedback.
Theme 2
Questions related to the syllabus for each topic, such as integrated pharmacotherapy of respiratory tract infections, were included.

Theme 3
Questions related to the preparation of lecture notes related to each topic, such as integrated pharmacotherapy of respiratory tract infections, were included.

Theme 4
Questions related to the preparation of examinations with model answers related to each topic, such as integrated pharmacotherapy of respiratory tract infections, were included.

Data Analysis
The performance of the ChatGPT model in providing answers for the education modules on integrated pharmacotherapy of infectious disease was extensively assessed. To ensure the robustness and credibility of the evaluation process, 3 highly qualified and experienced professors were carefully selected to assess the ChatGPT-generated answers. These professors have extensive knowledge and experience instructing modules on integrated pharmacotherapy of infectious diseases. Their extensive experience enables them to provide valuable insights and evaluations regarding the appropriateness, accuracy, and thoroughness of ChatGPT-generated responses. All 3 professors (one with a BPharm and PharmD from the United States; one with a BPharm, PharmD, and PhD in pharmacy practice from the United States; and one with a BPharm, MPharm, and PhD in clinical pharmacy from Malaysia) have more than 10 years’ experience in teaching modules on integrated pharmacotherapy of infectious disease in undergraduate and postgraduate programs.

A well-designed grading rubric was created to ensure consistency and justice in the evaluation procedure. This rubric served as a guide for professors to evaluate and grade ChatGPT’s responses. The evaluation rubric was meticulously crafted to include essential evaluation criteria, such as the relevance of the answers to the questions posed, their accuracy in reflecting the desired knowledge, and their comprehensiveness in addressing the specific aspects of the curriculum of the education modules on integrated pharmacotherapy of infectious disease. The professors meticulously scrutinized and evaluated the ChatGPT-generated responses, taking the established grading rubric into account. Their evaluations were based on their in-depth subject matter knowledge, pedagogical expertise, and curriculum development experience. The professors’ ratings were then averaged to guarantee a balanced and objective evaluation of the ChatGPT model’s performance.

In addition, the professors had the opportunity to provide qualitative comments and insights regarding the potential uses, benefits, and risks of using ChatGPT in the context of education modules on integrated pharmacotherapy of infectious disease. These additional qualitative contributions provide a deeper understanding of the implications and practical considerations associated with integrating ChatGPT into educational practices.

Our data analysis provides a rigorous and thorough examination of the performance of the ChatGPT model in the context of education modules on integrated pharmacotherapy of infectious disease by involving 3 accomplished professors, using a well-designed marking rubric, and incorporating qualitative insights. This meticulous methodology ensures the reliability and validity of the findings, allowing educators and researchers to make well-informed decisions regarding the implementation and potential benefits of ChatGPT in medical education.

Ethical Considerations
This project protocol was assessed and exempted for ethics approval by the Research Committee of the College of Medical Sciences, Azal University for Human Development (REC-2022-36).

Results

Theme 1: The Ability of ChatGPT to Design the Curriculum of Education Modules on Integrated Pharmacotherapy of Infectious Disease

Step 1: Problem Identification and General Needs Assessment

Overview
Our analysis of the experts’ opinions shows that ChatGPT was able to describe the need for the integrated pharmacotherapy curriculum in general for health care students and describe the issue of antibiotic resistance; however, it was unable to describe the importance of integrated pharmacotherapy of infectious disease. In general, the average of experts’ ratings of appropriateness and accuracy was 65%.

Potential Benefits
ChatGPT can help medical and health sciences educators by highlighting the importance of integrated pharmacotherapy curricula from reviewing the literature.

Potential Risks
ChatGPT could not describe the problem and carry out a general needs assessment for a specific population.

Recommendations
Medical and health sciences educators can use ChatGPT as a guide for understanding what is reported in the literature; then, they should be able to understand the problem and carry out a general needs assessment in the context of their countries with other methods.

Step 2: Targeted Needs Assessment

Overview
Our analysis of the experts’ opinions shows that ChatGPT was able to design a general initial questionnaire to use for the feasibility study of integrated pharmacotherapy; however, ChatGPT was unable to design a specific questionnaire related to integrated pharmacotherapy of infectious disease. Furthermore, ChatGPT was not able to design a qualitative study. The average of experts’ ratings of appropriateness and accuracy was 50%.
Potential Benefits
ChatGPT can help medical and health sciences educators to design a quick questionnaire to be used for conducting feasibility studies.

Potential Risks
There are many steps involved in designing valid and reliable questionnaires or qualitative interviews, which ChatGPT will not be able to undertake.

Recommendations
Medical and health sciences educators cannot use ChatGPT to develop valid and reliable questionnaires and qualitative interviews.

Step 3: Goals and Objectives
Overview
Our analysis of the experts’ opinions shows that ChatGPT could design the goals for the curriculum of the education modules on integrated pharmacotherapy of infectious disease, and the average of experts’ ratings of appropriateness and accuracy was 92%. ChatGPT could design general objectives for the curriculum of the education modules on integrated pharmacotherapy of infectious disease, and the average of experts’ ratings of appropriateness and accuracy was 80%.

Potential Benefits
ChatGPT can help medical and health sciences educators to design goals and objectives for the curriculum of the education modules on integrated pharmacotherapy of infectious disease.

Potential Risks
The goals and objectives suggested by ChatGPT were not specific and could not cover all learning objectives or outcome domains.

Recommendations
Medical and health sciences educators can use ChatGPT as a guide for preparing goals and objectives related to the curriculum of education modules on integrated pharmacotherapy of infectious disease.

Step 4: Educational Strategies
Overview
Our analysis of experts’ opinions shows that ChatGPT could help in the development of educational strategies, and the average of the experts’ ratings of appropriateness and accuracy was 75%.

Potential Benefits
ChatGPT can help medical and health sciences educators to develop educational strategies.

Potential Risks
The educational strategies suggested by ChatGPT could not be completed.

Recommendations
Medical and health sciences educators can use ChatGPT as a guide to develop educational strategies related to the curriculum of education modules on integrated pharmacotherapy of infectious disease.

Step 5: Evaluation and Feedback
Overview
Our analysis of experts’ opinions shows that ChatGPT could help suggest suitable evaluation and feedback, and the average of the experts’ ratings of appropriateness and accuracy was 85%.

Potential Benefits
ChatGPT can help medical and health sciences educators with teaching and learning evaluation and feedback methods (for different courses and programs).

Potential Risks
The suggested evaluation and feedback methods by ChatGPT could not be completed.

Recommendations
Medical and health sciences educators can use ChatGPT as a guide in the evaluation and feedback related to the curriculum of education modules on integrated pharmacotherapy of infectious disease.

Theme 2: Questions Related to the Syllabus for Each Topic, Such as Integrated Pharmacotherapy of Respiratory Tract Infections
Overview
Our analysis of the experts’ opinions shows that ChatGPT could help in syllabus design, and the average of the experts’ ratings of appropriateness and accuracy was 70%. However, the syllabus was not complete in terms of learning objectives, topics, and educational resources.

Potential Benefits
ChatGPT can, with caution, help medical and health sciences educators to design lecture notes for the curriculum of education modules on integrated pharmacotherapy of infectious disease.

Potential Risks
The suggested lecture notes by ChatGPT could not be completed and missed many important issues.

Recommendations
Medical and health sciences educators can use ChatGPT as a guide in preparing the syllabus of the curriculum of integrated pharmacotherapy of infectious disease.

Theme 3: Questions Related to the Preparation of Lecture Notes Related to Each Topic, Such as Integrated Pharmacotherapy of Respiratory Tract Infections
Overview
Our analysis of experts’ opinions shows that ChatGPT could help prepare lecture notes; however, the lecture notes were not complete, and the suggested learning objectives or outcomes for each lecture were not complete. The average of the experts’ ratings of appropriateness and accuracy was 65%.
**Potential Benefits**
ChatGPT can, with caution, help medical and health sciences educators to design the syllabus of the curriculum of integrated pharmacotherapy of infectious disease.

**Potential Risks**
The syllabus suggested by ChatGPT could not be completed and missed many important issues.

**Recommendations**
Medical and health sciences educators can use ChatGPT as a guide in preparing lecture notes for the curriculum of integrated pharmacotherapy of infectious disease.

**Theme 4: Questions Related to the Preparation of Examinations With Model Answers Related to Each Topic, Such as Integrated Pharmacotherapy of Respiratory Tract Infections**

**Overview**
Our analysis of expert’s opinions shows that ChatGPT could help in preparing model answers for examinations. However, the examinations did not cover all the learning objectives or outcomes. The average of experts’ ratings of appropriateness and accuracy was 70%.

**Potential Benefits**
ChatGPT can, with caution, help medical and health sciences educators to prepare model answers for different types of examinations related to the curriculum of integrated pharmacotherapy of infectious disease.

**Potential Risks**
The examination questions suggested by ChatGPT could not be completed and did not cover the learning objectives or outcomes.

**Recommendations**
Medical and health sciences educators can use ChatGPT as a guide in preparing examinations for the curriculum of integrated pharmacotherapy of infectious disease.

**Discussion**

**Background**
This study explored the ability of ChatGPT to help medical and health sciences educators in curriculum design, syllabus design, lecture notes preparation, and examination preparation. The findings of this study can be classified into 3 themes.

**Theme 1: Potential Benefits of Using ChatGPT in the Curriculum of Integrated Pharmacotherapy of Infectious Disease**
Our findings show that ChatGPT was able to help medical and health sciences educators, especially new educators, in all aspects of curriculum development with caution, and the experts rated the curriculum development aspects between 50% in the targeted needs assessment and 92% for suggestions about goals. Therefore, medical and health sciences educators can use ChatGPT as a guide in developing such a curriculum. ChatGPT is still in the early phase of use by educators worldwide, and it may be better in the near future to generate all steps related to such a curriculum appropriately and completely.

**Theme 2: Potential Risks of Using ChatGPT in the Curriculum of Integrated Pharmacotherapy of Infectious Disease**
Our findings show that there are potential risks associated with using ChatGPT in the development of the curriculum of integrated pharmacotherapy of infectious disease, syllabus design, lecture notes preparation, and examination preparation, such as missing important learning objectives or outcomes, various examination questions, and others. There are many limitations of ChatGPT; therefore, medical and health sciences educators should be aware of these limitations and use ChatGPT with caution, only as a guide to help them, and not rely 100% on it to do all work.

**Theme 3: Recommendations for Using ChatGPT in the Curriculum of Integrated Pharmacotherapy of Infectious Disease**
ChatGPT can help medical and health sciences educators in many ways, and they can use ChatGPT as a guide in curriculum design, syllabus design, lecture notes preparation, and examination preparation.

**Limitations**
A limitation of our study is that our methodology could benefit from additional clarification and elucidation, particularly in regard to the rating process and performance evaluation. Lack of explicit details regarding the specific criteria and scoring system used by evaluators to evaluate ChatGPT responses is another limitation. In the absence of a well-defined and standardized rating framework, subjectivity and potential ambiguity may be introduced into the evaluation process. This could impact the results’ dependability and comparability.

Another limitation is the reliance on qualitative assessments instead of quantitative measures for a more generalizable performance evaluation. The absence of quantitative metrics hinders the ability to objectively measure the system's accuracy, response time, and user satisfaction ratings, even though qualitative insights from educators provide valuable insights. Consequently, our findings may have limited applicability.

To address these limitations, future research could focus on developing a more exhaustive and standard rating framework and scoring system, and elucidating the reviewers’ criteria. Incorporating quantitative measures alongside qualitative assessments would provide a more robust and trustworthy evaluation of the performance of ChatGPT.

**Conclusions**
This study highlights the immense potential of ChatGPT as a valuable tool for medical and health sciences educators in various aspects of the curriculum of integrated pharmacotherapy of infectious disease. The findings emphasize both the benefits and risks of incorporating ChatGPT into educational practices, providing valuable insights for educators seeking to leverage...
AI technology to improve teaching and learning. This study demonstrates that ChatGPT can serve as a reliable resource for educators, especially those new to the field, in curriculum development, syllabus design, lecture note preparation, and examination preparation. Educators should exercise caution and use ChatGPT as a supplementary resource, rather than relying solely on its outputs, in order to ensure its effective and responsible use. Participating in workshops on AI technologies and ChatGPT can help educators to gain a deeper understanding of its capabilities and limitations, enabling them to make informed decisions and implement best practices.

**Authors’ Contributions**

YMAW conceptualized the study. AH and KWG carried out the formal analysis and acquired the funding. YMAW designed the methodology. YMAW and LCM were in charge of the study’s administration. KWG and CST were responsible for the software. YMAW supervised the study. AH and LCW were responsible for validation. YMAW drafted the manuscript. AH, KWG, CST, and LCM reviewed and edited the manuscript.

**Conflicts of Interest**

None declared.

**References**


**Abbreviations**

AI: artificial intelligence
LLM: large language model
©Yaser Mohammed Al-Worafi, Khang Wen Goh, Andi Hermansyah, Ching Siang Tan, Long Chiau Ming. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology–Head and Neck Surgery Certification Examinations: Performance Study

Cai Long, MD, MASc; Kayle Lowe, BSc; Jessica Zhang, BSc; André dos Santos, PhD; Alaa Alanazi, MD; Daniel O'Brien, MD; Erin D Wright, MDCM, MEd; David Cote, MPH, MD

1Division of Otolaryngology–Head and Neck Surgery, University of Alberta, Edmonton, AB, Canada
2Faculty of Medicine, University of Alberta, Edmonton, AB, Canada
3Alberta Machine Intelligence Institute, Edmonton, AB, Canada
4Department of Surgery, Creighton University, Omaha, NE, United States

Corresponding Author:
Cai Long, MD, MASc
Division of Otolaryngology–Head and Neck Surgery
University of Alberta
8440-112 Street
Edmonton, AB, T6G 2B7
Canada
Phone: 1 (780) 407 8822
Email: cai.long.med@gmail.com

Abstract

Background: ChatGPT is among the most popular large language models (LLMs), exhibiting proficiency in various standardized tests, including multiple-choice medical board examinations. However, its performance on otolaryngology–head and neck surgery (OHNS) certification examinations and open-ended medical board certification examinations has not been reported.

Objective: We aimed to evaluate the performance of ChatGPT on OHNS board examinations and propose a novel method to assess an AI model’s performance on open-ended medical board examination questions.

Methods: Twenty-one open-ended questions were adopted from the Royal College of Physicians and Surgeons of Canada’s sample examination to query ChatGPT on April 11, 2023, with and without prompts. A new model, named Concordance, Validity, Safety, Competency (CVSC), was developed to evaluate its performance.

Results: In an open-ended question assessment, ChatGPT achieved a passing mark (an average of 75% across 3 trials) in the attempts and demonstrated higher accuracy with prompts. The model demonstrated high concordance (92.06%) and satisfactory validity. While demonstrating considerable consistency in regenerating answers, it often provided only partially correct responses. Notably, concerning features such as hallucinations and self-conflicting answers were observed.

Conclusions: ChatGPT achieved a passing score in the sample examination and demonstrated the potential to pass the OHNS certification examination of the Royal College of Physicians and Surgeons of Canada. Some concerns remain due to its hallucinations, which could pose risks to patient safety. Further adjustments are necessary to yield safer and more accurate answers for clinical implementation.

(JMIR Med Educ 2024;10:e49970) doi:10.2196/49970

KEYWORDS
medical licensing; otolaryngology; otology; laryngology; ear; nose; throat; ENT; surgery; surgical; exam; exams; response; responses; answer; answers; chatbot; chatbots; examination; examinations; medical education; otolaryngology/head and neck surgery; OHNS; artificial intelligence; AI; ChatGPT; medical examination; large language models; language model; LLM; LLMs; wide range information; patient safety; clinical implementation; safety; machine learning; NLP; natural language processing
Introduction

The latest surge in artificial intelligence (AI) has been the development of ChatGPT by OpenAI as a large language model (LLM) trained on internet text data. LLMs have demonstrated remarkable capabilities in interpreting and generating sequences across various domains, including medicine. Since its initial release in November 2022, ChatGPT has been tested in various fields and corresponding standardized tests from high school to the postgraduate level for science, business, and law. The latest version of ChatGPT, based on GPT-4, was launched on March 14, 2023, with video and image input and is available to the public for a fee through the Plus and Enterprise services. In May and June 2023, iOS and Android apps, respectively, were made publicly available with added voice input capabilities. Image generation ability was added to ChatGPT using DALL-E 3 in October 2023 but remains restricted to Plus and Enterprise users. As of March 2023, GPT-4 has passed a diverse list of standardized examinations, including the Uniform Bar Examination, the SAT (Scholastic Assessment Test), Graduate Record Examinations (GRE), Advanced Placement (AP) examinations, and more [1]. In the field of medicine, ChatGPT has passed the United States Medical Licensing Examination (USMLE) and Medical College Admission Test (MCAT) [2,3]. Reviews on the application of ChatGPT in health care have been hopeful that it enhances efficiency, enables personalized learning, and encourages critical thinking skills among users, but concerns persist with the current limitations of ChatGPT’s knowledge, accuracy, and biases [4,5].

Concerns regarding misinformation were echoed when ChatGPT was tested against the US National Comprehensive Cancer Network (NCCN) guidelines for cancer treatment recommendations and found to be generally unreliable [6]. Its performance in fields such as ophthalmology, pathology, neurosurgery, cardiology, and neurology has been evaluated as being passable or near-passable [7-12]. Specifically, for surgical specialties, it was tested on multiple choice questions from the Ophthalmic Knowledge Assessment Program (OKAP) examination and both the oral and written board examinations for the American Board of Neurological Surgery (ABNS). For pathology and neurology, ChatGPT was presented with scenarios generated by experts in the respective fields and evaluated for accuracy [8,11]. When presented with 96 clinical vignettes encompassing emergency care, critical care, and palliative medicine, ChatGPT gave answers of variable content and quality. However, 97% of responses were deemed by physician evaluators as appropriate with no clinical guideline violations [13]. ChatGPT has also been tested for its performance on the tasks of medical note-taking and answering consultations [2,14]. To the best of our knowledge, ChatGPT or similar LLMs have not been evaluated for their performance in otolaryngology/head and neck surgery (OHNS).

In medical education, ChatGPT shows potential to generate quiz questions, reasonably explain concepts, summarize articles, and potentially supplement small group–based discussion by providing personalized explanations for case presentations [12,15]. Potential concerns include the generation of incorrect answers and false academic references [15].

There is a wide gap between competency on proficiency examinations or other medical benchmarks and the successful clinical use of LLMs. Appropriate use of well-calibrated output could facilitate patient care and increase efficiency. We present the first evaluation of an LLM (GPT-4) on the otolaryngology/head and neck surgery certification examination of the Royal College of Physicians and Surgeons of Canada (RCPSC) and propose a novel method to assess AI performance on open-ended medical examination questions.

The RCPSC is the accreditation and certifying agency that grants certifications to physicians practicing in medical and surgical specialties in Canada. The RCPSC examination is a high-stakes, 2-step comprehensive assessment comprising a written and applied component. To pass, candidates must achieve a score of 70% or higher on both components. The examination uses an open-ended, short-answer question format scored by markers using lists of model answers [16].

This research will provide valuable insights into the strengths and limitations of LLMs in medical contexts. The findings may inform the development of specialty-specific knowledge domains for medical education, enhance clinical decision-making by integrating LLMs into practice, and inspire further exploration of AI applications across industries, ultimately contributing to better health care outcomes and more effective use of AI technology in the medical field [17].

Methods

Twenty-one publicly available sample questions with model answers were obtained from the RCPSC website, which requires a login and is not indexed by Google. Random spot checks were performed to ensure that the content was not indexed on the internet. This was done by searching the question itself on Google and reading through the first 2 pages of results. Spot checks were done with every fifth question listed. Sample questions used were from previous official examinations. These questions can be found in Multimedia Appendix 1. Our assessment focuses on the text-only version of the model, referred to as GPT-4 (no vision) by OpenAI [18]. These questions were queried against GPT-4. A new chat session was initiated in ChatGPT for each entry to reduce memory retention bias, except for follow-up questions. Follow-up questions were asked in the same chat session. For example, a question with 2 follow-up questions would be repeated. Answers were recorded on April 11, 2023. To evaluate the effectiveness of prompting, questions were given with lead-ins prior to the first question in each scenario (“This is a question from an otolaryngology head and neck surgery licensing exam”), allowing the AI to generate answers that are more OHNS–specific. As LLMs lack fact-checking abilities, the consistency of answers is particularly important. To further assess consistency, each answer was regenerated twice and scored independently.

The answers were assessed and scored based on a newly proposed Concordance, Validity, Safety, Competency (CVSC) model (Table 1). Two physicians (CL and AA) independently scored the answers, and major discrepancies between the 2 scorers were sent to a third physician (DC) for a final decision. The maximum score was 34.
In the pursuit of a comprehensive understanding of its performance, we designed a new analytical framework. It drew inspiration from ACI (accuracy, concordance, and insight), a tool used by Kung et al [2] in evaluating the USMLE and many other multiple-choice medical board examinations.

Our assessment tool, the CVSC model, was developed based on several established assessment tools [2,18]. It provides an in-depth evaluation of answers generated by ChatGPT in terms of their concordance and homogeneity. Additionally, it scrutinizes the validity of the responses to identify hallucinations, which are a major concern in the application of LLMs in health care. Notably, it introduces a mechanism to report and flag responses that could potentially lead to unsafe or harmful practices for patients.

**Table 1.** Adjudication criteria for the Concordance, Validity, Safety, Competency model scoring system designed to assess the performance of ChatGPT in open-ended clinical questions.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Question</th>
<th>Answer options</th>
</tr>
</thead>
</table>
| Concordance    | Can the AI\(^a\) interpret the question correctly, and can it form a cohesive answer? | • Concordant: Explanation affirms the answer or is directly related to the question.  
• Discordant: Any part of the explanation contradicts itself or is not directly related to the question. |
| Validity       | Of all the information presented by AI, how much of it is valid (ie, a widely accepted opinion, an OHNS\(^b\) consensus, evidence based, scientifically proven, or the opinion of OHNS senior staff)? | • Valid: All information given is valid.  
• Mostly valid: ≥ 1 piece of information is invalid, but more than 50% of information presented is valid.  
• Fair: Proportion of valid to invalid information is equal.  
• Mostly invalid: ≥ 1 piece of information is valid, but more than 50% of information presented is invalid.  
• Invalid: All information is invalid.  
• Indeterminate: Contains generic information not applicable to given clinical scenario/question or does not answer the question.  
• Safety: If any option is chosen other than “valid” for the above question, answer the following question: “Does the answer contain information that is misleading, faulty, or nonproven according to the guidance, and if so, might it compromise patient safety? If the answer is yes, provide detail in the comment box below.” |
| Competency     | Regarding the overall performance of the AI, does it miss any important parts of the answer? | Numeric score that changes with each question. The value of the question is assigned according to an answer key based on the importance of the topic. |

\(^a\)AI: artificial intelligence.  
\(^b\)OHNS: otolaryngology–head and neck surgery.

**Results**

The preliminary data with questions and responses can be found in Multimedia Appendices 2-4.

For direct inquiries made to ChatGPT, the system achieved a cumulative score of 23.5 out of a possible 34, equaling 69.1%. The minimum passing score for the RCPSC examination is 70%. Further queries were conducted with ChatGPT with prompts explicitly indicating the focus to be OHNS specific. Under these conditions, as shown in Figure 1, ChatGPT exhibited superior performance, achieving a score of 75% (25.5/34) on the initial trial. When comparing the first attempt and the second attempt of ChatGPT, the first attempt was slightly better than the second attempt. The accuracy rate was found to be 72% (24.5/34) when the program was asked to regenerate its answers. However, the second set of answers demonstrated increased validity but less concordance.

This development marks a significant stride toward addressing patient safety concerns in using LLMs in health care. To our knowledge, the CVSC model is the first of its kind designed to systematically evaluate LLMs with a strong emphasis on patient safety.

Preliminary data were collected using Google Sheets and an ANOVA was performed using Excel (2022 version; Microsoft). This study only used publicly available information and did not involve humans, animals, or any of their information. Therefore, approval by the University of Alberta Research Ethics Board was not required.

The bulk of generated responses were found to be directly related to the question, with a concordance rate of 95%. Outliers in this instance were characterized by 2 divergent responses that were either self-contradictory or incongruous with the posed question. Figure 2 shows the validity of the answer groups. Overall, the majority (42/63, 67%) of responses were deemed valid, corroborated by either broadly accepted facts, OHNS consensus, evidence-based data, scientific validation, or alignment with the opinions of OHNS senior staff. A subset of the responses (17/63, 27%) contained partially invalid answers, with a minute fraction (2/63, 3%) being deemed mostly invalid. It was observed that these statements lacked scientific validity, adherence to evidence-based principles, or acceptance by the OHNS community; that is, they were what is known as hallucinations. There were some answers (2/63, 3%) that were verbose but did not contain information that could be assessed objectively.
To evaluate if there were any significant differences among the different groups, we performed an ANOVA using Microsoft Excel. We found there were no significant differences among the different groups ($F=0.06$, $F_{crit}=3.15$; $P=.93$).

**Figure 1.** Scoring details of 3 different groups of queries. A1: without prompt; A2: first attempt with prompt; A2b: second attempt with prompt.

**Figure 2.** Validity of different groups of queries. A1: without prompt; A2: first attempt with prompt; A2b: second attempt with prompt.

**Discussion**

**Principal Results**

The data presented in this study represent the first assessment of an LLM such as ChatGPT for OHNS specialty board examinations. It is also the first assessment of a medical specialty board examination with open-ended questions. The questions are in alignment with the RCPSC certifying examination for OHNS. This methodology is congruent with that used by the board examinations in Canada and several other nations.

This study used an official sample examination, which was meticulously reviewed by educational leads within the specialty and provides a strong correlation with real examination materials and difficulty level. Consequently, this assessment offers superior benchmarking capabilities, providing an authentic representation of the examination scores.
The open-ended questions endeavor to mimic real-life clinical scenarios, where physicians are frequently confronted with open-ended questions, challenging their capacity to reason and draw conclusions. Most other evaluations of the performance of LLMs such as ChatGPT are based on multiple-choice questions, showcasing AI’s ability to identify and incorporate key topics and crucial information. However, this format falls short in assessing the breadth of knowledge and reasoning capabilities of AI.

This research offers an initial exploration into these scenarios, providing a novel contribution to the ongoing discussion on how to accurately assess the capabilities of LLM systems such as ChatGPT in medical applications. By taking this approach, our study sets the stage for more thorough and nuanced evaluations of AI performance in settings that more closely resemble their real-world applications.

**The Concordance of Answers Generated by ChatGPT**

Overall, ChatGPT demonstrated considerable concordance; that is, its explanations affirmed the answer or were directly related to the question. Conversely, a response was deemed as discordant when any segment of the explanation contradicted itself or was not directly related to the question. This element of our assessment tool is particularly useful for LLMs such as ChatGPT, which are known to generate large amounts of text data with low information density.

During the evaluation, it was observed that the answers provided by ChatGPT were generally concordant (58/63, 92%) and directly addressed the question posed. Only 8% (5/63) of the responses contained conflicting or unrelated information. For instance, in 1 answer, ChatGPT incorrectly stated that the symptoms were solely caused by a bacterial infection, providing a lengthy explanation. However, in a subsequent explanation, it correctly identified the disease as juvenile recurrent parotitis with an unknown etiology, mentioning possible causes, such as autoimmune factors, obstruction, and infection, among others.

In another response, the initial part of the answer indicated that the frontal sinus bone was thicker than the adjacent bones, while the latter part stated that it was thinner. This conflicting information demonstrates the lack of inherent understanding of the text by ChatGPT, despite its self-generation of answers.

**The Validity of Answers Generated by ChatGPT**

The majority of the answers provided by ChatGPT were found to be valid: 67% (42/63) were identified as valid, 24% (15/63) were identified as mostly valid, and 10% (6/63) were found to be indeterminate, fair, or mostly invalid.

LLMs, including ChatGPT, have been known to generate hallucinations, which are characterized by blatant factual errors, significant omissions, and erroneous information generation [19]. The high linguistic fluency of LLMs allows them to interweave inaccurate or unfounded opinions with accurate information, making it challenging to identify such hallucinations.

For example, in one of the answers, ChatGPT introduced the term “recurrent bacterial parotitis,” which is not a recognized diagnosis accepted by the OHNS community. Similarly, in another response, ChatGPT mentioned “digital palpation” as one of the methods to identify the border of the frontal sinus. This method is a fabrication on the part of ChatGPT and is not recognized in established medical practice.

Overall, we observed that ChatGPT demonstrated high performance regarding foundational anatomy and the pathophysiology of OHNS disease presentations. In questions related to these topics, the answers generally received high validity scores, and fewer instances of hallucinations were observed. It is possible that the extensive text data available on these subjects allowed the LLM to draw more information and generate more accurate responses.

**Patient Safety Concerns in the Answers**

Hallucinations may present benign or harmful misinformation, with significant implications in the field of medicine. Such hallucinations could include misleading or incorrect data, and if followed by clinical practitioners, this may pose substantial risks to patient safety. In our evaluation, we asked evaluators to identify and red-flag any such statements they encountered.

Certain hallucinations, although inaccurate, do not critically impact patient safety. For instance, ChatGPT occasionally uses very outdated terminology. An example of this is the usage of “recurrent parotitis” rather than the current widely accepted terms “juvenile recurrent parotitis” or “recurrent parotitis of childhood.”

However, there are situations where ChatGPT’s inaccuracies could potentially compromise patient safety. For instance, when asked about the planes of a bicoronal approach for an osteoplastic flap, ChatGPT provided incorrect information, which could, in certain cases, jeopardize the flap. Similarly, ChatGPT suggested pharyngeal dilation as a surgical intervention in a scenario where it was not indicated. This could place a patient at risk of undergoing an unnecessary surgical procedure if the recommendation were followed precisely. Another instance of potentially harmful misinformation was ChatGPT’s suggestion of laryngotracheal reconstruction for an anterior glottic web, an approach that is excessively radical for the condition.

**The Overall Accuracy of the Results**

In our study, ChatGPT performed well and secured passing scores in all 3 tests: the unprompted test, the first attempt with a prompt, and the regenerated answer with a prompt, scoring 69%, 75%, and 72%, respectively.

It was noted that the AI performed very well on questions that require a specific knowledge base, such as anatomy- and physiology-related questions and disease diagnosis questions. Without prompting, the AI was found to generate more generalized responses that often lacked the depth and breadth typically expected in an OHNS board examination answer.

ChatGPT demonstrated potential in successfully navigating complex surgical specialty board examinations, specifically when presented with open-ended questions. Despite some observed discordance, the bulk of the information provided by the AI was clinically valid. Such features may prove highly...
beneficial for medical education, such as in equitable access to resources, particularly in low-resource settings where access to such information may not be readily available. The application of LLMs in medical education may also include writing examination questions, being an added “blind” marker, or even acting as a “bot examiner.” In addition, ChatGPT passing this examination may have implications on the format of the examination itself. Examination adjudicators and creators may have to consider alternative examination methods, including a shift toward oral-only examinations, to preserve the academic integrity of the RCPSC examinations.

Some inaccuracies identified were due to the use of outdated data. The AI’s text-prediction model may not frequently encounter updated information on the internet, leading to this issue. However, time-variant data present a challenge for LLMs due to their inability to differentiate between outdated data and newly published data supported by evidence. There is a lack of studies exploring the critical appraisal skills of LLMs, which are essential for clinical decision support.

Future work will investigate if domain-specific versions of GPT could offer increased accuracy and exhibit fewer hallucinations, thereby potentially reducing patient safety concerns. With the launch of ChatGPT Vision, subsequent studies could directly evaluate its interpretative ability for medical imaging in otolaryngology or other medical fields.

Limitations
While this study presents valuable insights into the performance of ChatGPT in open-ended OHNS questions, its inherent limitations must also be acknowledged. First, image-based questions could not be used for assessment due to the limitations of the currently available version of ChatGPT, which is based on GPT-4; the public version did not support visual data queries at the time of our test. Given that OHNS is a surgical specialty, key aspects such as surgical planning, anatomical identification, pathology recognition, and interpretation of intraoperative findings heavily depend on image analysis. Future versions of LLMs may be capable of handling such data, and we aspire to evaluate their efficacy in doing so. Second, the study’s data collection and validation methods require a more extensive set of questions. Only 21 questions were adopted from the RCPSC’s sample set for this study. For a more robust prediction and performance assessment, a larger question set is necessary. Third, we used prompt engineering to find appropriate prompts for the study; however, due to time and resource constraints, it is possible that other prompts may have allowed ChatGPT to achieve better results.

Conclusions
We evaluated the performance of ChatGPT by using it on a sample board-certifying examination of the RCPSC for OHNS, using our novel CVSC framework. ChatGPT achieved a passing score on the test, indicating its potential competence in this specialized field. Nevertheless, we have certain reservations, notably relating to the potential risk to patient safety due to hallucinations. Furthermore, the verbosity of the responses can compromise the practical application of LLMs. A systematic review done on ChatGPT’s performance on medical tests suggested that AI models trained on specific medical input may perform better on relevant clinical evaluations [20]. The development of a domain-specific LLM might be a promising solution to address these issues.

Acknowledgments
We thank Neil Saduka (Reeder AI) and Deepak Subburam (Copula AI) for their assistance and contributions during the course of this research.

Authors' Contributions
CL carried out the study design, data collection, and data analysis and drafted the manuscript. KL participated in data collection and data analysis. AdS participated in the study design. JZ participated in drafting the manuscript. AA helped with data collection. DO and EDW contributed to the final manuscript. DC participated in data collection, analysis, and reviewing and editing the manuscript. All authors reviewed and approved the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Sample questions from past examinations of the Royal College of Physicians and Surgeons. [DOCX File, 11 KB - mededu_v10i1e49970_app1.docx ]

Multimedia Appendix 2
Questions and ChatGPT answers (A1). [DOCX File, 914 KB - mededu_v10i1e49970_app2.docx ]

Multimedia Appendix 3
Questions and ChatGPT answers (A2a).
Multimedia Appendix 4
Questions and ChatGPT answers (A2b).

References
1. Varanasi L. AI models like ChatGPT and GPT-4 are acing everything from the bar exam to AP Biology. Here’s a list of difficult exams both AI versions have passed. Business Insider. 2023 Mar 21. URL: https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1 [accessed 2023-05-24]
18. GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses. OpenAI. URL: https://openai.com/product/gpt-4 [accessed 2023-11-01]

Abbreviations

ABNS: American Board of Neurological Surgery  
AI: artificial intelligence  
AP: Advanced Placement  
CVSC: Concordance, Validity, Safety, Competency  
GRE: Graduate Record Examinations  
LLM: large language model  
MCAT: Medical College Admission Test  
NCCN: National Comprehensive Cancer Network  
OHNS: otolaryngology/head and neck surgery  
OKAP: Ophthalmic Knowledge Assessment Program  
RCPSC: Royal College of Physicians and Surgeons of Canada  
SAT: Scholastic Assessment Test  
USMLE: United States Medical Licensing Examination
Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project

Nicholas I-Hsien Kuo1*, PhD; Oscar Perez-Concha1*, PhD; Mark Hanly1, PhD; Emmanuel Mnatzaganian2, MSc; Brandon Hao2, BA; Marcus Di Sipio2, BHSc; Guolin Yu2, BA; Jash Vanjara2, MD; Ivy Cerelia Valerie3, MD; Juliana de Oliveira Costa1, PhD; Timothy Churches4,5, MBBS; Sanja Lujic1, PhD; Jo Hegarty6, BIT; Louisa Jorm1, PhD; Sebastiano Barbieri1, PhD

1Centre for Big Data Research in Health, The University of New South Wales, Sydney, Australia
2The University of New South Wales, Sydney, Australia
3Medicines Intelligence Research Program, School of Population Health, The University of New South Wales, Sydney, Australia
4School of Clinical Medicine, University of New South Wales, Sydney, Australia
5Ingham Institute of Applied Medical Research, Liverpool, Sydney, Australia
6Sydney Local Health District, Sydney, Australia
* these authors contributed equally

Corresponding Author:
Nicholas I-Hsien Kuo, PhD
Centre for Big Data Research in Health
The University of New South Wales
Level 2, AGSM Building (G27), Botany St, Kensington NSW
Sydney, 2052
Australia
Phone: 61 0293850645
Email: n.kuo@unsw.edu.au

Abstract

Large-scale medical data sets are vital for hands-on education in health data science but are often inaccessible due to privacy concerns. Addressing this gap, we developed the Health Gym project, a free and open-source platform designed to generate synthetic health data sets applicable to various areas of data science education, including machine learning, data visualization, and traditional statistical models. Initially, we generated 3 synthetic data sets for sepsis, acute hypotension, and antiretroviral therapy for HIV infection. This paper discusses the educational applications of Health Gym’s synthetic data sets. We illustrate this through their use in postgraduate health data science courses delivered by the University of New South Wales, Australia, and a Datathon event, involving academics, students, clinicians, and local health district professionals. We also include adaptable worked examples using our synthetic data sets, designed to enrich hands-on tutorial and workshop experiences. Although we highlight the potential of these data sets in advancing data science education and health care artificial intelligence, we also emphasize the need for continued research into the inherent limitations of synthetic data.

(JMIR Med Educ 2024;10:e51388) doi:10.2196/51388

KEYWORDS

medical education; generative model; generative adversarial networks; privacy; antiretroviral therapy (ART); human immunodeficiency virus (HIV); data science; educational purposes; accessibility; data privacy; data sets; sepsis; hypotension; HIV; science education; health care AI

Introduction

Clinical data gathered from health care institutions are crucial for enhancing health care quality [1-3]. These data sets can feed into artificial intelligence (AI) and machine learning (ML) models to refine patient prognosis [4,5], diagnosis [6,7], and treatment optimization [8]. Furthermore, statistical models applied to these data sets can uncover association and causal paths [9]. However, stringent privacy regulations protecting patient confidentiality often hamper the prompt availability of these data sets for research and educational usage [10-14].
Gaining access to clinical and health care data sets is a critical aspect of health data science education. This exposure provides trainees with invaluable practical experience, offering profound insights into the complexities of real-world health care scenarios [15]. However, obtaining access to these sensitive data sets is a challenging endeavor—often involving a lengthy process of securing ethics approvals, institutional support, and data clearance [16]. Moreover, the approved users may be required to work on-site under the direct supervision of the data custodian to prevent data leakage [17]. These rigorous security measures, while essential for patient confidentiality, can hamper scalable training of future health data scientists.

During this era of big data, with a soaring demand for skilled health data scientists [18,19], synthetic data sets can bridge the gap between analytical skills and health context comprehension. As Kolaczyk et al [20] astutely asserted, “Theory informs principle, and principle informs practice; practice, in turn, informs theory.”

A promising solution to the lack of clinical and health care data is the utilization of generative AI to generate synthetic data sets. These data sets provide controlled, context-specific learning experiences that parallel real-world situations while maintaining patient privacy. The Health Gym project exemplifies this approach [21]. Leveraging generative adversarial networks (GANs) [22-24], Health Gym creates synthetic medical data sets, establishing a secure yet realistic platform for trainees to hone their health data analytical skills. The data sets, covering key health conditions such as sepsis, acute hypotension, and antiretroviral therapy (ART) for HIV infection, can be accessed at [25]. The project’s open-source code is also available on GitHub at [26] under the MIT License [27].

As an integral part of the Master of Science in Health Data Science Program at the University of New South Wales (UNSW), Australia [28] and a Datathon event [29], the Health Gym synthetic data sets have proven their versatility and effectiveness in enriching health care education. They are freely accessible to the wider research and education community while complying with stringent security standards such as those specified by Health Canada [30] and the European Medicines Agency [31], thus minimizing patient data disclosure risks.

In this viewpoint paper, we discuss the application of Health Gym synthetic data sets, their role in health data science education, and their potential in nurturing proficient health data scientists. We provide adaptable worked examples (accessible through Section A in Multimedia Appendix 1) by using our synthetic data sets, crafted to enrich hands-on tutorial and workshop experiences. We underline the importance of acknowledging the limitations of synthetic data to ensure their valid use in the creation of statistical models and AI applications in health care and the enhancement of health care education. Although synthetic data sets cannot supersede real-world data, they are a vital tool for training future health data scientists and supporting data-driven innovative approaches in health care.

**Ethics Approval**

We applied GANs to longitudinal data extracted from the MIMIC-III (Medical Information Mart for Intensive Care) [32] and the EuResist [33] databases to generate our synthetic data sets. This study was approved by the UNSW’s human research ethics committee (application HC210661). For patients in MIMIC-III, requirement for individual consent was waived because the project did not impact clinical care and all protected health information was deidentified [32]. For people in the EuResist integrated database, all data providers obtained informed consent for the execution of retrospective studies and inclusion in merged cohorts [34].

**Health Gym**

The currently available synthetic data sets for the Health Gym project were derived from MIMIC-III [32] and EuResist [33] databases. MIMIC-III is a comprehensive database of anonymized health data associated with patients admitted to the critical care units of the Beth Israel Deaconess Medical Center, including data on laboratory tests, procedures, and medications. The EuResist network aims to develop a decision support system to optimize ART for individuals living with HIV, leveraging extensive clinical and virological data.

After applying published selection or exclusion criteria, we extracted relevant data from databases that could facilitate the development of patient care algorithms. These data sets, focusing on sepsis, acute hypotension, and ART for HIV, served as the basis for our synthetic data creation. The synthetic data generation employed in the Health Gym was accomplished using GANs. The GAN model, as shown in Figure 1, consists of 2 primary components: a generator and a discriminator.

The process starts by sampling real patient records (depicted in pink) and employing the generator to create synthetic patient records (depicted in violet). Both the real and synthetic records are then forwarded to the discriminator network, which is tasked with differentiating the genuine data from the counterfeit. Both networks are trained in an adversarial process—the generator is updated to create more realistic records, while the discriminator is refined to identify generated records more accurately. As a result, the quality of the synthetic data is progressively enhanced, and the synthetic patient records become increasingly representative of the ground truth. The iterative training concludes when the discriminator can no longer reliably distinguish the synthetic records from the real records. Refer to more details in Kuo et al [21].

Leveraging generative AI, Health Gym provides highly authentic clinical data sets, enriching health care education. Each data set undergoes rigorous quality assessment and security verification (detailed in Section B of Multimedia Appendix 1). These synthetic data sets foster engaging learning experiences, aiding educators in developing tailored educational strategies. The following sections will illuminate the application of Health Gym in university-level courses, exemplified through ART for HIV data set.

---

**References**

2. MIMIC-III (Medical Information Mart for Intensive Care) [32].
3. EuResist integrated database [33].
4. Ethics Approval [34].

---

**Author Details**

Kuo et al

**Institutional Affiliation**

University of New South Wales (UNSW), Australia

**Contact**

CH Kuo (corresponding author)

chkuo@unsw.edu.au
Synthetic ART for HIV Data Set

The Health Gym data sets contain mixed-type longitudinal data, including numerical, binary, and categorical variables. They encompass patient demographics, vital signs measurements, and pathology results. The data sets hence reflect the complexities of real-life data, thereby making them suitable for training health data scientists in university courses. This paper will primarily delve into the application of synthetic data in health care education focusing on the ART for HIV data set. Readers interested in the sepsis and the acute hypotension data sets should refer to Section C in Multimedia Appendix 1.

Data Set Description

Our synthetic HIV data set, informed by the selection or exclusion criteria proposed by Parbhoo et al [35] and drawn from the EuResist database, targets individuals living with HIV who initiated therapy after 2015 per the World Health Organization’s guidelines [36]. ART for HIV typically includes a mix of 3 or more antiretroviral agents from at least 2 distinct medication classes. The dynamism of ART lies in its frequent regimen modifications resulting from various circumstances such as treatment failure due to poor adherence or viral resistance, intolerance to ART, clinical events such as pregnancy or coinfections, or optimization of therapy to support better adherence, reduce drug-drug interactions, maximize ART response, or prevent the emergence of drug-resistant viral strains [36,37].

In addition to ART information, the data set encompasses vital indicators of ART success and disease progression, namely, viral load (VL) and CD4 cell count. Successful ART is often indicated by VL below 1000 copies/mL, while a CD4 cell count exceeding 500 cells/mm³ signifies healthy immunological status [36]. The complex interactions of these elements in our data set create a rich learning platform for health data science education.

Table 1 encapsulates the data set’s 3 numeric, 5 binary, and 5 categorical variables. Numeric variables include VL, CD4 cell count, and relative CD4 laboratory test results. Treatment regimens follow those of Tang et al [38], breaking down the ART regimen into several parts. The data set includes 50 combinations of 21 unique medications. The antiretroviral medication classes are nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), nonnucleoside reverse transcriptase inhibitors (NNRTIs), integrase inhibitors (INIs), protease inhibitors (PIs), and pharmacokinetic enhancers (pk-En). We deconstructed the ART regimen into its constituent parts: base drug combination (base drug combo), complimentary INIs (comp INIs), comp NNRTIs, extra PIs, and extra pk-En. The base drug combo primarily consists of NRTIs, with inclusion of other antiretroviral classes as well.

Recognizing the notable amount of missing data in the original EuResist database, we added a suffix (M) to variables to denote whether measurements were recorded at specific time points. In the authentic data set, measurements were reported at 24.27% (129,835/534,960) for VL (measured), 22.21% (118,815/534,960) for CD4 (measured), and 85.13% (455,411/534,960) for drug (measured). The absence of some CD4 and VL records may be attributable to specific clinical practices and the frequency of test requests [39-42]. For instance, it is common for clinicians to discontinue requesting a CD4 cell count if the previous result exceeded 500 cells/mm³ and the individual had an undetectable VL. Similarly, VL is typically measured in the first 3 months, at 6 months, 12 months, and then annually.

Constructed using the GAN model developed by Kuo et al [43], this data set comprises 8916 synthetic patients tracked over 60 months, resulting in 534,960 records (8916 × 60). Figure 2 showcases a sample generated by the code in Figure 3 [44,45]. Each record features 15 columns, including a patient identifier, a time point, and 13 ARTs for HIV variables highlighted in Table 1. The synthetic data sets can be freely accessed in [46] and [47] on Figshare, a digital platform for research output sharing.
<table>
<thead>
<tr>
<th>Variable name</th>
<th>Data type</th>
<th>Unit</th>
<th>Valid categorical options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viral load (VL)</td>
<td>numeric</td>
<td>copies/mL</td>
<td>N/A</td>
</tr>
<tr>
<td>Absolute count for CD4 (CD4)</td>
<td>numeric</td>
<td>cells/µL</td>
<td>N/A</td>
</tr>
<tr>
<td>Relative count for CD4 (Rel CD4)</td>
<td>numeric</td>
<td>cells/µL</td>
<td>N/A</td>
</tr>
<tr>
<td>Gender</td>
<td>binary</td>
<td>N/A</td>
<td>Male, Female</td>
</tr>
<tr>
<td>Ethnicity (Ethnic)</td>
<td>categorical</td>
<td>N/A</td>
<td>Asian, African, Caucasian, other</td>
</tr>
<tr>
<td>Base drug combination (Base drug combo)</td>
<td>categorical</td>
<td>N/A</td>
<td>FTC\textsuperscript{b} + TDF\textsuperscript{c}, 3TC\textsuperscript{d} + ABC\textsuperscript{e}, FTC + TAF\textsuperscript{f}, DRV\textsuperscript{g} + FTC + TDF, FTC + RTVB\textsuperscript{h} + TDF, other</td>
</tr>
<tr>
<td>Complementary integrase inhibitor (Comp INI)</td>
<td>categorical</td>
<td>N/A</td>
<td>DTG\textsuperscript{i}, RAL\textsuperscript{j}, EVG\textsuperscript{k}, not applied</td>
</tr>
<tr>
<td>Complementary nonnucleoside reverse transcriptase inhibitor (Comp NNRTI)</td>
<td>categorical</td>
<td>N/A</td>
<td>NVP\textsuperscript{l}, EFV\textsuperscript{m}, RPV\textsuperscript{n}, not applied</td>
</tr>
<tr>
<td>Extra protease inhibitor (Extra PI)</td>
<td>categorical</td>
<td>N/A</td>
<td>DRV, RTVB, LPV\textsuperscript{o}, RTV\textsuperscript{p}, ATV\textsuperscript{q}, not applied</td>
</tr>
<tr>
<td>Extra pharmacokinetic enhancer (Extra pk-En)</td>
<td>binary</td>
<td>N/A</td>
<td>False, True</td>
</tr>
<tr>
<td>Viral load measured (VL) (M)\textsuperscript{r}</td>
<td>binary</td>
<td>N/A</td>
<td>False, True</td>
</tr>
<tr>
<td>CD4 (M)</td>
<td>binary</td>
<td>N/A</td>
<td>False, True</td>
</tr>
<tr>
<td>Drug recorded (M)</td>
<td>binary</td>
<td>N/A</td>
<td>False, True</td>
</tr>
</tbody>
</table>

\textsuperscript{a}N/A: not applicable.
\textsuperscript{b}FTC: emtricitabine.
\textsuperscript{c}TDF: tenofovir disoproxil fumarate.
\textsuperscript{d}3TC: lamivudine.
\textsuperscript{e}ABC: abacavir.
\textsuperscript{f}ABC: abacavir.
\textsuperscript{g}TAF: tenofovir alafenamide.
\textsuperscript{h}DRV: darunavir.
\textsuperscript{i}RTVB: ritonavir.
\textsuperscript{j}DTG: dolutegravir.
\textsuperscript{k}RAL: raltegravir.
\textsuperscript{l}EVG: elvitegravir.
\textsuperscript{m}NVP: nevirapine.
\textsuperscript{n}EFV: efavirenz.
\textsuperscript{o}RPV: rilpivirine.
\textsuperscript{p}LPV: lopinavir.
\textsuperscript{q}RTV: ritonavir.
\textsuperscript{r}(M): measured.
Figure 2. Inspecting the antiretroviral therapy for an HIV data set (output of the code in Figure 3).

```python
# The top 5 rows of the ART for HIV dataset
VL CD4 Rel CD4 Gender Ethnic Base Drug Combo
0 29.682271 79.42838 38.81405 1 0 0
1 29.223980 657.47808 38.35538 3 0 0
2 28.728864 605.13987 58.68520 2 0 0
3 18.150875 390.96809 58.20856 1 0 0
4 23.818307 641.77024 29.34472 1 0 0

Comp. INI Comp. NNRTI Extra PI Extra pk-En VL (M) CD4 (M) Drug (M)
0 0 0 3.0 3.0 0.0 1.0 1.0
1 0 0 1.0 5.0 0.0 0.0 1.0
2 0 0 3.0 5.0 0.0 0.0 1.0
3 0 0 3.0 5.0 0.0 0.0 1.0
4 0 0 3.0 5.0 0.0 0.0 1.0

PatientID Timestep
0 0
1 0
2 0
3 0
4 0

---
Shape of the dataset (#WxH, #I)
---
Column names
["VL", "CD4", "Rel CD4", "Gender", "Ethnic", "Base Drug Combo", 
"Comp. INI", "Comp. NNRTI", "Extra PI", "Extra pk-En", "VL (M)", 
"CD4 (M)", "Drug (M)", "PatientID", "Timestep"],
class="object"
---
Total number of synthetic patients 8956
---

# The top 5 rows of data relating to synthetic patient no. 100
VL CD4 Rel CD4 Gender Ethnic Base Drug Combo
6000 15960.138 2531.7308 23.75599 2.8 3.0 0.0
6001 14500.320 654.7245 21.43551 2.8 3.0 0.0
6002 12901.162 634.8021 26.45708 2.8 3.0 0.0
6003 25388.082 553.4155 85.61571 2.8 3.0 0.0
6004 53873.276 1295.7946 77.82381 3.0 3.0 0.0

Comp. INI Comp. NNRTI Extra PI Extra pk-En VL (M) CD4 (M) Drug (M)
6000 0.0 0.0 1.0 0.0 1.0 0.0 1.0
6001 0.0 0.0 1.0 0.0 1.0 0.0 1.0
6002 0.0 0.0 1.0 0.0 1.0 0.0 1.0
6003 0.0 0.0 1.0 0.0 1.0 0.0 1.0
6004 0.0 0.0 1.0 0.0 1.0 0.0 1.0

Drug (M) PatientID Timestep
6000 0.0 0
6001 0.0 1
6002 0.0 100 1
6003 0.0 100 3
6004 0.0 100 4
```

Figure 3. Code in Python for generating the output shown in Figure 2. This code uses pandas [44] and NumPy [45]. Base drug combo: base drug combination; comp INI: complementary integrase inhibitor; comp NNRTI: complementary nonnucleoside reverse transcriptase inhibitor; PI: protease inhibitor; pk-En: pharmacokinetic enhancer; VL: viral load.

Applications and Case Studies

This section highlights the use of our synthetic ART for HIV data set in a collaborative Datathon event and as an effective teaching tool at UNSW for medical education.

Center for Big Data Research in Health Data Science Datathon

The synthetic data set for ART for HIV was a central component of the UNSW Center for Big Data Research in Health Datathon [48], an event merging theoretical learning with practical application. The Datathon was an enriching exercise in multidisciplinary collaboration. The event involved 6 teams, with a total of 24 participants, offering a tangible experience in
data analysis. The student teams were supported by a group of mentors—a blend of data scientists, clinicians, health professionals, and government health informatics specialists from a local health district in Sydney, Australia [49]. The data scientists and the panel of authors of the Health Gym project (ie, Kuo et al [21]) elaborated on the technical aspects and navigated the participants through the intricacies of data analysis, including the assumptions we made to use the data (eg, time 0 corresponded to the date of ART initiation, the laboratory tests occurred before modifications in therapy). Meanwhile, clinicians and health professionals provided their expertise to guide students toward meaningful research questions (eg, discussing VL and CD4 count monitoring, drug-drug interactions, and metabolic toxicity [50]). Government health informaticians, experienced in electronic medical records and real-world population health application and impact, evaluated the usefulness of the students’ findings.

This collaborative effort facilitated a comprehensive learning experience, encompassing the development of analytical models, data visualization, and effective communication of research outcomes. Using our synthetic data sets, participants gained valuable insights into working with data sets that emulate real-world health scenarios, thereby providing a bridge between theoretical academia and practical execution.

We summarize the findings of the 2 participating teams below. Detailed reports for Team 1 and Team 2 can be found in Section D and Section E of Multimedia Appendix 1, respectively. In addition, the associated codes for the 2 teams can be found in Section A of Multimedia Appendix 1.

Findings of Team 1

Team 1 investigated the effectiveness of medications, categorized by antiretroviral class, in achieving HIV suppression. Utilizing survival analysis, they assessed the time between the initiation of ART to the first occurrence of viral suppression, defined as VL below 1000 copies/mL [36]. They also assessed the time to CD4 cell count exceeding 500 cells/mm$^3$ [51], which indicates a healthy immunological status.

With Cox proportional hazards models [52] featuring time-varying covariates, the team identified particular antiretroviral agents associated with viral suppression. These findings were purely associative due to data set limitations, which did not account for factors such as age, socioeconomic status, comorbidities, and concurrent medications (of other illnesses).

Findings of Team 2

Team 2 focused on predicting the necessity of altering an individual’s ART regimen over a 5-year time span, factoring in disease flare-ups, resistance, or side effects. They formulated a “sliding search” function that generated individual records for each 12-month period, with predictions for antiretroviral modification and adherence to therapy in the subsequent year by using neural networks. The team’s methodology produced promising results, with an accuracy rate of 78% in predicting antiretroviral modification and 93% in predicting adherence to therapy. The algorithm detected trends in CD4 and VL results across the 12-month periods, which appeared to be the key predictive features. In addition, the team suggested that there could be potential benefits from exploring recurrent neural networks (eg, long short-term memory [53]).

Serving as UNSW Coursework Materials

Beyond their utility in the Datathon, our synthetic data sets contribute to UNSW courses in the Master of Science in Health Data Science Program [54], namely, HDAT9800 Visualization & Communication and HDAT9510 Machine Learning II.

HDAT9800 teaches future health data scientists the skills to visually communicate complex data effectively to diverse audiences. The course emphasizes the significance of clear data visualization and advocates for transparency and reproducibility in scientific work. It employs R [55] and Python [56] to demonstrate best practices in data analysis and visualization. Our synthetic data sets provide rich resources to enhance the learning in this setting. For instance, Marchesi et al [57] used our data sets to present patient states via t-distributed stochastic neighbor embedding visualization techniques [58].

Meanwhile, HDAT9510 explores advanced modern ML algorithms and methods such as convolutional neural networks [59], autoencoders [60], and reinforcement learning (RL) [61]. As the synthetic data sets consist of time-series variables, students can develop both feedforward and recurrent neural networks. See example models built using our data set in Marchesi et al [57] with recurrent neural networks and even decision trees [62] and hidden Markov models [63], as in a similar data set suggested by Wu et al [64]. Furthermore, with the presence of nonnumeric variables, students can learn about embedding [65]—transforming nonnumeric levels into real-valued vectors so that similar levels that are closer in the vector space carry more analogous meaning. The presence of missing data in the synthetic data sets also encourages students to formulate plausible assumptions about the structure of the clinical data set prior to data modelling.

We provide 3 adaptable worked examples using our ART for HIV data set, suitable for workshops and lectures. The associated codes for the worked examples can be found in Section A of Multimedia Appendix 1. Our synthetic data set supports a variety of student engagements, from understanding complex data structures to developing advanced RL algorithms for optimizing clinical interventions. Moreover, the low patient disclosure risk associated with our data sets (refer to Section B in Multimedia Appendix 1) eliminates the need for ethics approval [66]. This makes these data sets ideal for a range of settings—from small seminars to larger lecture groups.

Worked Example 1

The first exercise, focused on data visualization using Python, compares VL trends over time among patients who commenced their ART with different base drug combos, against the general trend in all patients. The results of our worked example are depicted in Figure 4.

This multifaceted exercise requires students to create sub–data sets based on specific starting base drug combos (ie, FTC + TDF [emtricitabine + tenofovir disoproxil fumarate] and 3TC + ABC [lamivudine + abacavir]), extract data for defined
periods, and familiarize themselves with box and violin plots [67]. They are also tasked with organizing the visual data as side-by-side plots.

Through this exercise, students will understand the limitations of box plots, which cannot visualize underlying data distributions. They will learn about the additional insights provided by advanced plotting techniques such as violin plots. In addition, students will note that people who start with FTC + TDF and those who start with 3TC + ABC display similar patterns as the overall ART for HIV cohort. The overlap of the interquartile ranges across all box plots indicates a consistent behavior.

**Figure 4.** Viral load distribution. Subplot (A) shows a box plot comparison of viral load across base drug combinations across time, and subplot (B) shows a violin plot comparison of viral load across base drug combinations across time. Grey indicates all patients, red indicates those initiating treatment with FTC + TDF (emtricitabine + tenofovir disoproxil fumarate), and blue indicates those initiating treatment with 3TC + ABC (lamivudine + abacavir). VL: viral load.

**Worked Example 2**

The second exercise delves into survival analysis using R [55], building on insights from the initial data visualization task. The exercise continues to compare results among people starting with the base drug combo of FTC + TDF and those initiating with the base drug combo of 3TC + ABC. The goal is to estimate the time necessary for a person on ART to successfully suppress their VL. The results of our worked example are depicted in Figure 5.

This task proves to be more complex than the first, requiring HIV domain knowledge, such as an understanding that a reasonable threshold for ART in HIV treatment is 1000 copies/mL [36]. This threshold indicates slowed viral replication and immune system damage. Thus, students should select patients who commence ART with VL above 1000 copies/mL (ie, not experiencing the outcome of interest at baseline).

Creating an appropriate data set for survival analysis is key, as is pinpointing when each patient’s VL first drops to or below 1000 copies/mL. In addition, students need to grasp the concept of right censoring [68] and utilize Kaplan-Meier curves [69] for time-to-event estimations. This offers an opportunity to engage with the influential survival package [70] in the R language. Upon examining the results in Figure 5, students will note no significant differences in the timing of VL suppression between people who started with the base drug combo of FTC + TDF and those who initiated with the base drug combo of 3TC + ABC.
Worked Example 3

The third exercise immerses students in the process of developing an RL agent using Python. RL is a type of ML that learns an evidence-based policy to connect states (the current scenario) to actions (the potential responses to that scenario). In the context of our HIV treatment example, states refer to the representation of the patient’s current health status and medication history, while action refers to the selection of medication to use in response to each state.

The RL agent selects an action based on a policy that optimizes for maximum cumulative rewards, even as environments evolve. This approach has particular relevance to health care. Clinicians often need to adapt treatment plans to each patient’s unique circumstances, and RL can help them to individualize treatment durations, dosages, or types. For example, they may alter the regimen, class, or specific agents of medication to better serve the patient’s needs. The outcomes of our example are visualized in Figure 6. This exercise highlights the potential of RL to enhance patient care through personalization—an aspect that is becoming increasingly important in today’s medical landscape.

This complex exercise is designed for advanced students, posing challenges across multiple dimensions. It commences with data wrangling, where students scrutinize numeric variable distributions and evaluate the necessity for transformations such as rescaling, normalization [71], power transformation [72], or Box-Cox transformation [73].

In the next stage, students encounter categorical feature representation for medication regimens, practicing their skills in implementing embeddings. Advanced students can explore transfer learning for feature representation [74]. This exercise also presents real-world challenges, requiring students to handle mixed-type data progression. During the model fitting phase, students must employ suitable ML models, distinguishing between RL method archetypes [75] and considering their clinical implications.

Data visualization is the next task, encouraging students to articulate model-derived insights into digestible visuals for a diverse audience. The concluding phase involves refining assumptions and model performance, incorporating multiple tests to identify optimal hyperparameters [76]. Here, students peek into the “black box” nature of ML and gain an intuition for effective module combinations [77-79]. This step becomes critical for causal inference tasks that necessitate rigorous input data validation [80].

Figure 6 showcases the strategy employed by an RL agent in HIV therapy. Heatmaps visualize the relative frequencies of chosen actions (ie, the selected antiretroviral), where each tile represents a unique action and its frequency as a proportion of all actions. The example output shows that the RL agent consistently suggests the EFV + RAL (efavirenz + raltegravir)—a combination of comp NNRTIs and comp INIs—4.39% of the time, while never recommending the RPV + RAL (rilpivirine + raltegravir) combination. More information on the steps taken to create the output for this task can be found in Section F of Multimedia Appendix 1.
Discussion

This paper demonstrates the transformative potential of synthetic health data sets in health care education, especially in the evolving context of generative AI integration. These data sets provide a realistic representation of real-world health data complexities while preserving patient confidentiality, facilitating experiential learning, skills enhancement, and interdisciplinary collaboration. However, this significant stride toward AI integration in education is not without challenges, and the creation of AI models trained on curated quality data sets emerges as a promising research area.

Despite our best efforts, the Health Gym synthetic data sets might not fully capture the complexity and diversity of real-world scenarios. For instance, some critical health determinants such as socioeconomic status [81] and comorbidities [82] are missing from the ART for HIV synthetic data sets. The absence of these factors mirrors the broader issues concerning data accessibility [83], particularly when it involves specialized or rare disease information. Furthermore, synthetic data might overlook uncontrolled variables or confounders inherent in real-world data [84,85], posing pedagogical challenges. However, this limitation is not solely attributable to our methodology. Since the socioeconomic status variable is not present in the EuResist database, our model lacked the necessary reference data from the outset.

In the field of health data science, proficient data set management and curation are essential due to the decentralized nature of health care data collection. Many entities contribute to health data, each using their own systems [86]. Privacy laws such as Australia’s Privacy Act 1988 [87] and the United States’ Health Insurance Portability and Accountability Act [88] complicate the sharing of data, resulting in a fragmented view of patient information.

Record linkage techniques [89] such as probabilistic matching [90] bridge this gap by linking disparate data records, offering a more comprehensive view of a patient’s health. Nevertheless, our synthetic data sets, despite their potential, carry limitations such as the absence of a master linkage key [91], thereby reducing their applicability in university courses for data management and curation. Having such linked data sets are also great for health data science students to test hypotheses on the effects of comorbidities. Our experiences from the Datathon suggest that the Health Gym synthetic data sets are best used for creating algorithms to enhance patient care within specific disease management paradigms.

Our Health Gym initiative leverages a unique application of generative AI, differing from those used in emerging AI-assisted chatbots, which have also shown promise as potent educational tools. AI chatbots, with their personalized and interactive responses using large language models, can significantly incite interest and foster self-directed learning in medical students [92]. However, advanced AI tools such as OpenAI’s ChatGPT [93] and Google’s BARD [94] bring with them valid concerns around precision, reliability, potential misuse, and adherence to academic integrity [95,96]. In contrast, the synthetic clinical data sets, the generative product of our Health Gym project, offer controlled, scenario-specific learning environments that
closely reflect real-world conditions while preserving patient privacy.

Access to clinical data sets is integral to health data science education, but the necessity of maintaining patient confidentiality can hinder the training of future health data scientists on a larger scale. This may exacerbate the digital divide [97,98], which is a prominent challenge in the broader AI integration into education. As we shift toward AI-driven educational resources, it is essential to prioritize equitable access across varied socioeconomic backgrounds. Future research should evaluate the long-term effects of AI on student learning, clinical judgment, patient outcomes, and the development of educational resources for effective AI integration. The secure, realistic synthetic data sets of Health Gym may provide a valuable solution, potentially facilitating equal access to educational materials.

Conclusion

Despite their limitations, the Health Gym synthetic health data sets have demonstrated their value in educating and training future health data scientists. Their integration into interdisciplinary platforms such as Datathon illustrates their potential in promoting collaborative learning, skills enhancement, and innovative research. In addition, synthetic data sets offer a learning platform that balances realistic health scenario representation with data privacy preservation.

Although we have primarily demonstrated the utility of Health Gym’s synthetic data sets by using the ART for HIV data set, we emphasize the importance of the additional acute hypotension and sepsis data sets that we have developed (see Section C in multimedia Appendix 1). These data sets broaden the scope of medical education by providing insight into managing illnesses in intensive care units, encompassing a unique set of measurements and pathology information. As such, these synthetic data sets offer students an enriched, realistic learning environment for health data science education, complementing the HIV data set and furthering the applicability and versatility of synthetic health data.

The majority of generative ML research is centered on computer vision [99,100] and, to a lesser extent, natural language processing [101], leaving clinical health care data relatively unexplored. This gap suggests a valuable opportunity for future research, particularly considering that clinical data being longitudinal, mixed-type time series variables have a fundamentally different nature. As demonstrated in our prior studies [21,43,102], we have ascertained that our synthetic data sets attain a robust level of validity and are readily available to support both clinical research and medical pedagogy: predictive models instantiated on our synthetic data sets parallel those of the original data sets in their characteristics. We will focus our future work on comparing synthetic data sets created using various generative ML architectures, for example, GANs, variational autoencoders [103], diffusion probabilistic models [102,104], and transformer-based models [105].

GANs, like other ML models, can only optimize according to predefined optimization functions. Given the current lack of research on the use of GANs in health care, more utility studies are necessary to fully comprehend the potential of our synthetic data sets. We are committed to continuing collaboration with clinicians and health professionals to better understand the practical strengths and weaknesses of synthetic data sets, including how to better evaluate and contain the risk of private information disclosure. Through these collective efforts, we aim to improve the quality of synthetic data sets, enhancing hands-on learning experiences for students in health data analytics.

Acknowledgments

This study benefited from data provided by the EuResist Network EIDB, and this project has been funded by a Wellcome Trust Open Research Fund (reference 219691/Z/19/Z). JdOC is supported by the Medicines Intelligence Center of Research Excellence (grant 1196900).

Authors' Contributions

Authors NI-HK and SB were responsible for the design, implementation, and validation of the deep learning models employed to generate the synthetic data sets for the Health Gym project. The inception of Datathon was conceived by OP-C and MH who liaised with various disciplinary personnel to realize this initiative. JdOC contributed specialist knowledge on antiretroviral therapy for HIV to Datathon, while JH offered expertise in the evaluation of Datathon projects. Furthermore, TC and SL, alongside OP-C and MH, leveraged their extensive teaching experience to guide Datathon participants and explore further applications of the Health Gym synthetic data sets. LJ provided key insights on the potential risk of sensitive information disclosure. Datathon participants EM, BH, MDS, GY, JV, and ICV gave critical feedback on the strengths and shortcomings of the synthetic data sets, in addition to providing valuable reflections on the event itself. This manuscript was prepared by NI-HK. All authors contributed to interpreting the findings and revising the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary data.
References


28. Graduate certificate in Health Data Science, The University of New South Wales. URL: https://www.unsw.edu.au/study/postgraduate/graduate-certificate-in-health-data-science?studentType=Domestic [accessed 2023-12-27]


47. Kuo NIH. The Health Gym v2.0 synthetic antiretroviral therapy (ART) for HIV dataset. Figshare. URL: https://figshare.com/articles/dataset/The_Health_Gym_v2_0_Synthetic_Antiretroviral_Therapy_ART_for_HIV_Dataset/22827878 [accessed 2023-12-27]
54. Master of Science in Health Data Science. The University of New South Wales. URL: https://www.unsw.edu.au/study/postgraduate/master-of-science?eq_plac=&studentType=Domestic [accessed 2023-12-27]
55. R: a language and environment for statistical computing. R-Project. URL: https://www.r-project.org/


86. Datasets - CHeReL. Centre for Health Record Linkage. URL: [https://www.cherel.org.au/datasets](https://www.cherel.org.au/datasets) [accessed 2023-12-27]


93. Kuo et alJMIR MEDICAL EDUCATION 2024 | vol. 10 | e51388 | p.235https://mededu.jmir.org/2024/1/e51388

94. Introducing ChatGPT. OpenAI. URL: [https://openai.com/blog/chatgpt](https://openai.com/blog/chatgpt) [accessed 2023-12-27]


Abbreviations

3TC: lamivudine
ABC: abacavir
AI: artificial intelligence
ART: antiretroviral therapy
Base drug combo: base drug combination
Comp INI: complementary integrase inhibitor
EFV: efavirenz
FTC: emtricitabine
GAN: generative adversarial network
INI: integrase inhibitor
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
NNRTI: nonnucleoside reverse transcriptase inhibitor
NRTI: nucleotide reverse transcriptase
PI: protease inhibitor
pk-En: pharmacokinetic enhancer
RAL: raltegravir
RL: reinforcement learning
RPV: rilpivirine
TDF: tenofovir disoproxil fumarate
UNSW: University of New South Wales
VL: viral load

Please cite as:


Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project

JMIR Med Educ 2024;10:e51388

URL: https://mededu.jmir.org/2024/1/e51388
doi:10.2196/51388
PMID:38227356
Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study

Firas Haddad¹, BSc; Joanna S Saade², MD

¹Faculty of Medicine, American University of Beirut, Beirut, Lebanon
²Department of Ophthalmology, American University of Beirut Medical Center, Beirut, Lebanon

Corresponding Author:
Joanna S Saade, MD
Department of Ophthalmology
American University of Beirut Medical Center
Bliss Street
Beirut, 1107 2020
Lebanon
Phone: 961 1350000 ext 8031
Email: js62@aub.edu.lb

Abstract

Background: ChatGPT and language learning models have gained attention recently for their ability to answer questions on various examinations across various disciplines. The question of whether ChatGPT could be used to aid in medical education is yet to be answered, particularly in the field of ophthalmology.

Objective: The aim of this study is to assess the ability of ChatGPT-3.5 (GPT-3.5) and ChatGPT-4.0 (GPT-4.0) to answer ophthalmology-related questions across different levels of ophthalmology training.

Methods: Questions from the United States Medical Licensing Examination (USMLE) steps 1 (n=44), 2 (n=60), and 3 (n=28) were extracted from AMBOSS, and 248 questions (64 easy, 122 medium, and 62 difficult questions) were extracted from the book, Ophthalmology Board Review Q&A, for the Ophthalmic Knowledge Assessment Program and the Board of Ophthalmology (OB) Written Qualifying Examination (WQE). Questions were prompted identically and inputted to GPT-3.5 and GPT-4.0.

Results: GPT-3.5 achieved a total of 55% (n=210) of correct answers, while GPT-4.0 achieved a total of 70% (n=270) of correct answers. GPT-3.5 answered 75% (n=33) of questions correctly in USMLE step 1, 73.33% (n=44) in USMLE step 2, 60.71% (n=17) in USMLE step 3, and 46.77% (n=116) in the OB-WQE. GPT-4.0 answered 70.45% (n=31) of questions correctly in USMLE step 1, 90.32% (n=56) in USMLE step 2, 96.43% (n=27) in USMLE step 3, and 62.90% (n=156) in the OB-WQE. GPT-3.5 performed poorer as examination levels advanced (P<.001), while GPT-4.0 performed better on USMLE steps 2 and 3 and worse on USMLE step 1 and the OB-WQE (P<.001). The coefficient of correlation (r) between ChatGPT answering correctly and human users answering correctly was 0.21 (P=.01) for GPT-3.5 as compared to −0.31 (P<.001) for GPT-4.0. GPT-3.5 performed similarly across difficulty levels, while GPT-4.0 performed more poorly with an increase in the difficulty level. Both GPT models performed significantly better on certain topics than on others.

Conclusions: ChatGPT is far from being considered a part of mainstream medical education. Future models with higher accuracy are needed for the platform to be effective in medical education.

(JMIR Med Educ 2024;10:e50842) doi:10.2196/50842

KEYWORDS
ChatGPT; artificial intelligence; AI; board examinations; ophthalmology; testing
quite famous for its outstanding ability to answer questions and assist in many tasks [4].

Medical education relies highly on standardized multiple-choice examinations to test medical students in an objective and consistent way. Ophthalmologists in the United States pass through the United States Medical Licensing Examination (USMLE) steps 1, 2, and 3, the Ophthalmic Knowledge Assessment Program (OKAP), and the Board of Ophthalmology (OB) Written Qualifying Examination (WQE) by the time they become practicing physicians. Undergraduate and graduate medical students rely on different tools available to prepare for these examinations.

One limitation of the current tools for medical education is the lack of personalization. Question banks used today do not tailor their explanations to users; rather, they present one explanation for each question to all its users. ChatGPT and other LLMs, if proven to be accurate in their ability to answer questions, can provide robust explanations to users, and users can then ask specific questions they need further clarification on. This can be very helpful and educational for users as it can tailor to the needs of each user and help them fill specific knowledge gaps they may have. Additionally, the GPT-3.5 model is freely available to everyone, while GPT-4.0 is available at a premium. As such, it is essential to compare these models to assess whether GPT-4.0’s hypothetical increased abilities justify the price of the membership.

The question of how ChatGPT can be integrated for use in medical education has emerged. With the complexity of ophthalmology, the ability of ChatGPT to accurately answer ophthalmology questions could be of significant value to medical students and residents preparing for the USMLE, OKAP, and OB-WQE. It is also important to compare the performance of both GPT-4.0 and GPT-3.5, since GPT-4.0 is marketed as a more intelligent version of its predecessor.

Therefore, the aim of this study is to evaluate the performance of ChatGPT on ophthalmology questions from USMLE steps 1, 2, and 3, the OKAP, and the OB-WQE using both GPT-3.5 and GPT-4.0. We hypothesize that ChatGPT’s responses are comparable to those of human experts in the field, and that GPT-4.0 performs better than GPT-3.5. The results of this study could have implications for the future use of ChatGPT in medical education and training, and for the development of more efficient and effective tools for examination preparation.

### Methods

#### Data Sets

Different data sets were used for the different examinations due to the lack of a central service for all examinations. Questions that included pictures or tables were automatically excluded and were not queried on ChatGPT. AMBOSS [5], a question bank and popular resource for the USMLE was used for steps 1, 2, and 3. A total of 44 questions were included for step 1, 60 for step 2, and 28 for step 3. AMBOSS highlights the difficulty of each question and the percentage of people who chose each answer choice. This allowed us to compare the performance of ChatGPT to the general population [5]. For the OKAP and OB-WQE, 248 questions across the different chapters were taken from *Ophthalmology Board Review Q&A* by Glass et al [6].

#### Prompt Engineering

The style and the prompt of the questions asked to ChatGPT have been shown to have an impact on the answer given. To standardize the process of asking the questions to ChatGPT, questions were all formatted in the same way on Word (Microsoft Corp). After removing questions with pictures or tables, the questions were formatted in the manner described by Gilson et al [7]. The question stem was consolidated in 1 paragraph, and then each answer choice was placed on a separate line. Furthermore, the answer choices were separated by 2 empty lines from the main question stem; this was done to optimize the accuracy of the results, avoiding any effect the question format may have on ChatGPT’s ability. An example prompt is shown in Textbox 1.

**Textbox 1.** An example of a prompt (written by the authors).

**Question:** What medical discipline deals with conditions of the eye

A. Dermatology  
B. Endocrinology  
C. Ophthalmology  
D. Rheumatology

**Data Analysis**

Data analysis was conducted using both Python (Python Software Foundation) and Excel. Excel was used to determine the percentage of correct answers. Python (Python Anaconda Spyder 5.3.3) was used to determine the percentage of correct answers by difficulty, test type, and topic. A chi-square test was conducted on Python to determine whether there are any significant differences in answering correctly based on test type and difficulty. Python was also used to compute the coefficient of correlation (and $P$ value) between ChatGPT answering.
correctly and the percentage of users who got the correct answer. Point-biserial was used to compute the correlation between ChatGPT answering questions correctly and humans answering correctly. Other tests included chi-square analysis and the Fisher exact test to investigate relationships between 2 categorical variables (difficulty level, correct or incorrect answers, etc).

**Ethical Considerations**

Since this study does not involve any human participants, institutional review board approval is not necessary for the purpose of this study. This study also respects the rights and copyright of the owners of the resources used and has obtained their approval for using the questions without sharing the questions anywhere in the data or paper.

**Results**

A total of 380 questions were queried on ChatGPT. The number of questions for each examination were 44 for step 1, 60 for step 2, 28 for step 3, and 248 for the OKAP and OB-WQE. The total percentage of correct answers was 55% (n=210) across all examinations for GPT-3.5, while it was 70% (n=270) for GPT-4.0. Table 1 shows the number and percentage of correct answers for each examination by each GPT model.

Between GPT-3.5 and GPT-4.0, GPT-4.0 performed significantly better on USMLE steps 2 and 3 and the OB-WQE but not on USMLE step 1. While GPT-3.5’s performance decreased with an increase in the examination level \((P<.001)\), GPT-4.0 performed better on USMLE steps 2 and 3 and poorer on the OB-WQE and USMLE step 1. The coefficient of correlation \((r)\) between ChatGPT answering correctly and the percentage of humans answering correctly on AMBOSS was 0.21 \((P=.01)\) for GPT-3.5 and –0.31 \((P<.001)\) for GPT-4.0.

Table 2 highlights the percentage of correct questions based on the difficulty level in the AMBOSS questions and in the OB-WQE questions.

Table 3 highlights the performance of both models according to the different topics in the OB-WQE and OKAP questions. Performance for both models was nonrandom, with both models performing better on certain topics such as corneal diseases, pediatrics, retina, ocular oncology, and neuro-ophthalmology.

**Table 1.** Performance of GPT-3.5 and GPT-4.0 on various examinations.

<table>
<thead>
<tr>
<th>Examination</th>
<th>Correct answers provided by models(^a), n (%)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GPT-3.5</td>
<td>GPT-4.0</td>
</tr>
<tr>
<td>USMLE(^b) step 1</td>
<td>33 (75)</td>
<td>31 (70.45)</td>
</tr>
<tr>
<td>USMLE step 2</td>
<td>44 (73.33)</td>
<td>56 (90.32)</td>
</tr>
<tr>
<td>USMLE step 3</td>
<td>17 (60.71)</td>
<td>27 (96.43)</td>
</tr>
<tr>
<td>OB-WQE(^c)</td>
<td>116 (46.77)</td>
<td>156 (62.90)</td>
</tr>
</tbody>
</table>

\(^a\)\(P<.001\) for between-model comparisons in the proportion of correct answers.  
\(^b\)USMLE: United States Medical Licensing Examination.  
\(^c\)OB-WQE: Board of Ophthalmology Written Qualifying Examination.
Table 2. Performance of GPT-3.5 and GPT-4.0 according to different difficulty levels.

<table>
<thead>
<tr>
<th>Difficulty level</th>
<th>AMBOSS</th>
<th>GPT-4.0</th>
<th>GPT-3.5</th>
<th>AMBOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Board of Ophthalmology difficulty level</td>
<td>Correct answers a, n (%)</td>
<td>ChatGPT's performance (correct answers), n (%)</td>
</tr>
<tr>
<td>1</td>
<td>49 (76)</td>
<td>1</td>
<td>19 (100)</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>73 (59)</td>
<td>2</td>
<td>43 (91)</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>35 (56)</td>
<td>3</td>
<td>38 (84)</td>
<td>53</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
<td>10 (59)</td>
<td>37</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>5</td>
<td>4 (66.67)</td>
<td>26</td>
</tr>
</tbody>
</table>

a P = .04 on comparing the performance of GPT-4.0 across different difficulty levels.
b P = .003 on comparing the performance of GPT-4.0 across different difficulty levels.
c P = .49 on comparing the performance of GPT-3.5 across different difficulty levels.
d P = .18 on comparing the performance of GPT-3.5 across different difficulty levels.
e N/A: not applicable.

Table 3. Performance of GPT-3.5 and GPT-4.0 on various included topics.

<table>
<thead>
<tr>
<th>Category</th>
<th>Correct answers by GPT-4.0 a, n (%)</th>
<th>Topic</th>
<th>Correct answers by GPT-3.5 b, n (%)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornea, external disease, and anterior segment</td>
<td>28 (74)</td>
<td>Cornea, external disease, and anterior segment</td>
<td>25 (66)</td>
<td>.45</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>20 (61)</td>
<td>Glaucoma</td>
<td>16 (48)</td>
<td>.32</td>
</tr>
<tr>
<td>Lens and cataract</td>
<td>22 (88)</td>
<td>Lens and cataract</td>
<td>8 (32)</td>
<td>&lt;.001 c</td>
</tr>
<tr>
<td>Neuro-ophthalmology</td>
<td>15 (54)</td>
<td>Neuro-ophthalmology</td>
<td>16 (57)</td>
<td>.06</td>
</tr>
<tr>
<td>Oculofacial, plastics, and orbit</td>
<td>17 (50)</td>
<td>Oculofacial, plastics, and orbit</td>
<td>10 (29)</td>
<td>.08</td>
</tr>
<tr>
<td>Pediatric ophthalmology and strabismus</td>
<td>14 (61)</td>
<td>Pediatric ophthalmology and strabismus</td>
<td>9 (34)</td>
<td>.07</td>
</tr>
<tr>
<td>Refractive management and optics</td>
<td>17 (50)</td>
<td>Refractive management and optics</td>
<td>14 (41)</td>
<td>.46</td>
</tr>
<tr>
<td>Retina and ocular oncology</td>
<td>24 (73)</td>
<td>Retina and ocular oncology</td>
<td>18 (54)</td>
<td>.12</td>
</tr>
</tbody>
</table>

a P = .02 for differences in the number of correct answers provided by GPT-4.0 among different categories.
b P = .03 for differences in the number of correct answers provided by GPT-3.5 among different topics.
c Significant at P < .05.

Discussion

Principal Findings

Our results indicate that GPT-4.0 is superior to GPT-3.5, and that GPT-3.5 has a below-average accuracy in answering questions correctly. The total proportion of correct answers for GPT-3.5 was 55% (n=210), which is considered a poor performance, while that of GPT-4.0 was 70% (n=270), which is an almost average performance [7]. Students typically must achieve 59%-60% of correct answers to pass, and students perform with an average of around 70%-75% on the aforementioned board examinations [7]. It is interesting to note that GPT-3.5’s performance decreased as examination levels increased. This is probably due to the more clinical nature of the examinations. This was not the case for GPT-4.0, which performed best on USMLE steps 2 and 3.

This study investigates the correlation between ChatGPT-3.5 and -4.0 providing a correct answer and the percentage of human users who provided the answer correctly on AMBOSS. For GPT-3.5, a correlation coefficient of 0.21 (P=.01) was noted; whereas, this correlation coefficient was –0.31 (P<.001) for GPT-4.0. This implies that GPT-4.0 performed better on questions that fewer users answered correctly.

Although our study is limited in that it did not divide the questions into categories such as diagnosis, treatment, basic knowledge, or surgical planning questions. Looking closely at the lens and cataract section in which the model failed (32% of correct answers for GPT-3.5), it was noted that all the correct
answers were basic knowledge questions. Surprisingly, an analysis of incorrect answers showed that almost half of the incorrectly answered questions were also basic knowledge questions. For instance, in one of the questions, the model was unable to identify the collagen fiber type in cataract—a piece of information that is widely available on the internet.

On the other hand, GPT-4.0 performed significantly better on basic knowledge questions. One may postulate that since GPT-4.0 was fed a larger database than was GPT-3.5, it has better abilities in answering basic knowledge questions than GPT-3.5. A study by Taloni et al [8] also noted a significant difference in performance between the 2 models in the cataract and anterior segment diseases categories.

It is unclear why it performed so poorly in the lens and cataract section. It could be hypothesized that managing diseases of the lens and cataract may be mostly surgical. This may not have been fed into this language learning model. Furthermore, surgical management requires input from images and videos, which were excluded from our paper and may have caused the drastic difference in performance. Further studies with more questions are needed to answer this question.

Table 2 outlines the percentage of correct answers based on the difficulty level on both models. GPT-4.0 performed poorer on questions with greater difficulties on both AMBOSS and OB-WQE questions, whereas this observation was not significant in GPT-3.5, indicating that it performed almost equally well across difficulty levels. Gilson et al [7] also reported a similar finding for GPT-3.5. Further studies are needed to explain those findings.

This study also examined the proportion of correct answers based on the different topics. Both models performed significantly better on certain topics than others. This is a novel finding not reported in other studies assessing the performance of ChatGPT. It is interesting to further explore this association and why a model would perform on certain topics better than others. It could be hypothesized that questions on topics such as oculoplastic, which rely on surgical techniques and knowledge of aesthetics, may be more difficult for AI models compared to topics such as strabismus and pediatric ophthalmology. To our knowledge, our study is among the first few to assess the abilities of GPT-4.0 in medical examinations across various levels of education and various board examinations.

When reviewing the explanations provided by ChatGPT, it was noted that the model would randomly either explain the provided answer choice or not. It is particularly remarkable to read how it justified the wrong answer choices. More studies are needed to emphasize and assess the answer justifications of the model. Indeed, having solid explanations is essential for it to become a reliable medical education tool.

Our study is unique in that it assesses the capabilities of ChatGPT in answering ophthalmology-related questions in contrast to other studies that assessed its ability to succeed in general examinations such as USMLE steps 1 and 2. Furthermore, this is the first study to assess the ability of ChatGPT to answer questions of a certain discipline across all its examination levels. Finally, this is among the first studies to compare GPT-4.0’s performance to GPT-3.5’s performance in medical examinations.

ChatGPT can be a great add-on to mainstream resources to study for board examinations. There have been reports of using it to generate clinical vignettes and board examination–like questions, which can create more unique practice opportunities for students. Additionally, our study also assesses the accuracy of the 2 models on board examination questions related to ophthalmology. Students can input questions they need help with on the platform, and receive an answer and explanation by using the platform. If the student is not satisfied with the answer provided, or has further questions, he or she can respond to the model and receive a more personalized answer. This is crucial as it significantly decreases the time needed to study and also creates a tailored study experience for each student’s needs.

However, ChatGPT needs further optimization before it can be considered a mainstream tool for medical education. The image feature was not present in GPT-3.5 and was introduced in GPT-4.0. This feature is available only on demand and is yet to be available to all users. Its accuracy and reliability are yet to be established for examination purposes. Many questions were excluded due to them containing images, which is a
considerable limitation considering the visual nature of ophthalmology. Even in the text-only questions, ChatGPT had moderate accuracy in answering questions across different difficulties and levels. This study is, however, limited by the small number of questions, particularly in the USMLE steps, due to the absence of a large number of ophthalmology questions in the resources used to prepare for these examinations. More studies are needed, which input a larger number of questions. This study also does not assess the repeatability of ChatGPT's answers; however, a study by Antaki et al [13] reported near-perfect repeatability.

Conclusions
Overall, this study suggests that ChatGPT has moderate accuracy in answering questions. Its accuracy decreases in nature as the examinations become more advanced and more clinical in nature. In its current state, ChatGPT does not seem to be the ideal medium for medical education and preparation for board examinations. Future models with more robust capabilities may soon become part of mainstream medical education. More studies are needed, which input a larger number of questions to verify the results of this study and attempt to find explanations for many of the intriguing findings.

Acknowledgments
We thank AMBOSS and Thieme Publishers for granting access to the questions for use in this present study. All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Conflicts of Interest
None declared.

References
5. Powerful learning and clinical tools combined into one platform. AMBOSS. URL: https://www.amboss.com/ [accessed 2023-03-05]
Evaluation of ChatGPT’s Real-Life Implementation in Undergraduate Dental Education: Mixed Methods Study

Argyro Kavadella¹, DDS, MSc, PhD; Marco Antonio Dias da Silva², DDS, MSc, PhD; Eleftherios G Kaklamanos¹,³,⁴, DDS, MSc, PhD; Vasileios Stamatopoulos⁵, BSc, MSc; Kostis Giannakopoulos¹, DDS, PhD

¹School of Dentistry, European University Cyprus, Nicosia, Cyprus
²Research Group of Teleeducation and Teledentistry, Federal University of Campina Grande, Campina Grande, Brazil
³School of Dentistry, Aristotle University of Thessaloniki, Thessaloniki, Greece
⁴Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates
⁵Information Management Systems Institute, ATHENA Research and Innovation Center, Athens, Greece

Corresponding Author:
Argyro Kavadella, DDS, MSc, PhD
School of Dentistry
European University Cyprus
6, Diogenes street
Engomi
Nicosia, 2404
Cyprus
Phone: 357 22559620
Email: a.kavadella@euc.ac.cy

Abstract

Background: The recent artificial intelligence tool ChatGPT seems to offer a range of benefits in academic education while also raising concerns. Relevant literature encompasses issues of plagiarism and academic dishonesty, as well as pedagogy and educational affordances; yet, no real-life implementation of ChatGPT in the educational process has been reported to our knowledge so far.

Objective: This mixed methods study aimed to evaluate the implementation of ChatGPT in the educational process, both quantitatively and qualitatively.

Methods: In March 2023, a total of 77 second-year dental students of the European University Cyprus were divided into 2 groups and asked to compose a learning assignment on “Radiation Biology and Radiation Protection in the Dental Office,” working collaboratively in small subgroups, as part of the educational semester program of the Dentomaxillofacial Radiology module. Careful planning ensured a seamless integration of ChatGPT, addressing potential challenges. One group searched the internet for scientific resources to perform the task and the other group used ChatGPT for this purpose. Both groups developed a PowerPoint (Microsoft Corp) presentation based on their research and presented it in class. The ChatGPT group students additionally registered all interactions with the language model during the prompting process and evaluated the final outcome; they also answered an open-ended evaluation questionnaire, including questions on their learning experience. Finally, all students undertook a knowledge examination on the topic, and the grades between the 2 groups were compared statistically, whereas the free-text comments of the questionnaires were thematically analyzed.

Results: Out of the 77 students, 39 were assigned to the ChatGPT group and 38 to the literature research group. Seventy students undertook the multiple choice question knowledge examination, and examination grades ranged from 5 to 10 on the 0-10 grading scale. The Mann-Whitney U test showed that students of the ChatGPT group performed significantly better (P=.045) than students of the literature research group. The evaluation questionnaires revealed the benefits (human-like interface, immediate response, and wide knowledge base), the limitations (need for rephrasing the prompts to get a relevant answer, general content, false citations, and incapability to provide images or videos), and the prospects (in education, clinical practice, continuing education, and research) of ChatGPT.

Conclusions: Students using ChatGPT for their learning assignments performed significantly better in the knowledge examination than their fellow students who used the literature research methodology. Students adapted quickly to the technological environment of the language model, recognized its opportunities and limitations, and used it creatively and efficiently. Implications for practice:
the study underscores the adaptability of students to technological innovations including ChatGPT and its potential to enhance educational outcomes. Educators should consider integrating ChatGPT into curriculum design; awareness programs are warranted to educate both students and educators about the limitations of ChatGPT, encouraging critical engagement and responsible use.

**KEYWORDS**
ChatGPT; large language models; LLM; natural language processing; artificial Intelligence; dental education; higher education; learning assignments; dental students; AI pedagogy; dentistry; university

**Introduction**

**Background**

The emergence of ChatGPT (OpenAI) in November 2022 represents the third significant technological breakthrough in information technology impacting education, following the introduction of Web 2.0 over a decade ago [1] and e-learning’s surge during the COVID-19 pandemic [2]. ChatGPT is an artificial intelligence (AI) tool that offers benefits and opportunities in higher education including increased student engagement, collaboration, personalized feedback, and accessibility. However, it is characterized by a limited database, posing challenges such as the restricted ability to answer medical questions and the potential for inaccurate and biased responses. There are also concerns regarding legal and ethical implications, plagiarism, and academic integrity [3-5].

The research on AI and its implementation in academic education is a prominent subject; a Google Scholar search for “artificial intelligence and dental education,” yielded 100,000 results and approximately 18,000 results for “ChatGPT and higher education” (on June 9, 2023). AI technology has evolved to unprecedented levels, transforming professions, revolutionizing workflows, and reshaping human-machine interactions. ChatGPT, the most recent milestone in natural language processing AI models, has been enabling advanced conversational capabilities and expanding the boundaries of AI-powered communication. Interest in ChatGPT applications encompasses both clinical practice [6,7] and higher education [3,8-11], with promising results.

**Relevant Prior Research**

Within the higher education landscape, it has been suggested that dental curricula at universities need to be updated due to the AI paradigm shift [9,12,13]. This involves defining a fundamental dental curriculum for both undergraduate and postgraduate levels and establishing learning outcomes related to dental AI [8]. Cotton et al [3] and Halaweh [14] proposed strategies to ensure the ethical and responsible use of AI tools in higher education. Fergus et al [10] evaluated academic answers generated using ChatGPT, and Bearman et al [15] in their review on AI in higher education discussed the shifting dynamics of authority and the relationships among teachers, students, institutions, and technologies. Gimpel et al [16] in their extensive discussion paper proposed guidelines and recommendations for students and lecturers and urged the universities for a multistakeholder dialogue to implement efficient and responsible use of generative AI models in higher education.

Roganovic et al [17] performed a cross-sectional web-based survey among experienced dentists and final-year undergraduate students from the School of Dental Medicine, University of Belgrade, Serbia, to investigate their current perspectives and readiness to accept AI into practice. Responders, especially final-year students, showed a lack of knowledge regarding AI use in medicine and dentistry (only 7.9% of them were familiar with AI use) and were skeptical (only 34% of them believed that AI should be used in dental practice); the underlying reasons were fear of being replaced by AI, as well as a lack of regulatory policies, since students and—at a lesser degree—dentists were concerned that using AI could legally complicate the clinical practice [17].

Chan and Hu [11] reported different results in exploring students’ perceptions of generative AI and ChatGPT in teaching and learning through a web-based questionnaire; the study revealed a generally positive attitude toward generative AI, with students demonstrating a good understanding of this technology, its benefits, and limitations, despite its novel public appearance. Generative AI is a special category of AI designed to learn from the characteristics of its input and generate outputs with similar characteristics. In contrast to most AI models that perform specific tasks based on predefined rules and patterns, generative AI models use advanced algorithms to find the underlying patterns of the input data (eg, text, images, sounds, and videos) and “generate” entirely new content of the same type [11]. Students recognized the potential for personalized feedback and learning support, brainstorming, writing assistance, and research capabilities and stated they would integrate technologies like ChatGPT in their studies and future careers, but they were also concerned about becoming overreliant on them. They moreover expressed concerns about data accuracy, privacy, ethical issues, and the impact on personal development [11]. Students’ perceptions of the learning environment and the teaching strategies have a significant impact on their approach to learning and the learning outcomes (positive perceptions lead to a deep approach to learning), thus being of pedagogical interest to educators and institutions [11,18]. The influence of AI tools on students’ engagement and perceptions was investigated by Nazari et al [19]: they conducted a randomized controlled trial to examine the efficacy of an AI-powered writing tool (Grammarly) for postgraduate students and concluded that students in the intervention group demonstrated significant improvement in engagement (behavioral, emotional, and cognitive), self-efficacy, and academic emotions (positive and negative), domains that address learning behavior, which lead to self-development and underpin authentic pedagogy.
Aims of the Study

Despite numerous publications about AI and large language models (LLMs), the majority involve discussion papers, viewpoint articles, and positions [3,13,16,20,21], with few being exploratory, cross-sectional, or questionnaire-based studies [11,17,19]. To our knowledge, so far, no experimental studies have been identified, wherein ChatGPT was in vivo implemented by students within the teaching process, and the outcomes were comprehensively evaluated.

Therefore, this study aimed to address this gap by implementing ChatGPT within the learning process and conducting a quantitative (differences between examination grades) and qualitative (thematic analysis of the free-text comments of the evaluation questionnaire) evaluation of the outcomes (mixed methods research study).

Methods

Ethical Considerations

The study’s research protocol was reviewed and approved by the Vice-Rector for Research and External Affairs and the President of the Institutional Committee on Bioethics and Ethics of the European University Cyprus.

Study Design: Challenges

The study was conceptualized, organized, and refined in February 2023 and realized in March 2023. Of note is that ChatGPT appeared publicly on November 30, 2022; in March 2023, ChatGPT-3.5 was freely available (and was mostly used by the students), whereas ChatGPT-4 had just emerged (few students used this). The study was not a stand-alone research endeavor; instead, it constituted part of students’ educational activities embedded within the semester’s educational program. As this was the first attempt to implement ChatGPT in the educational process and there were no existing research studies in the literature to refer to, and adding to the limited knowledge on ChatGPT’s properties and limitations at the time, the authors encountered various challenges while organizing the research design. Therefore, to anticipate potential issues that could affect student learning or compromise the study’s outcomes, they conducted a systematic, forward-looking analysis of the research process, considering each step and taking proactive measures to mitigate any challenges or obstacles that may have arisen.

Study Design: Implementation

The second-year dental students (77 students) of the School of Dentistry, European University Cyprus were randomly divided into 2 large groups and were asked to compose an assignment on “Radiation Biology and Radiation Protection in the Dental Office.” The subject of Dentomaxillofacial Radiology is taught through theoretical lectures, laboratory training, and practical training during 2 semesters, and students’ learning assignments are embedded within the lectures’ program as an alternative to traditional lecturing. Student learning assignments to replace lectures followed by in-class presentation and discussion is a methodology used within the “Dentomaxillofacial Radiology” module whenever the topic is suitable for such an approach. Students usually work collaboratively to perform the assignments by searching the internet for scientific reliable sources and compiling the results into a PowerPoint slide presentation, including the references they used. Students of both groups were asked to work in small subgroups to compose the assignments, where each subgroup would comprise 3-7 students, decided among them. It is worth mentioning that the European University Cyprus School of Dentistry is an English-speaking School, educating students from over 30 countries encompassing different ethnic, educational, and cultural backgrounds; therefore, the study’s sample could be considered diverse.

One large group would compose the assignment through literature research (the traditional method for assignments) and the other group would use the ChatGPT tool for the assignment (pose prompts and register the answers), also submitting a slide presentation. Students were given 1 month to deliver the assignment, and they were informed that they would present their presentations in class on a designated day.

Moreover, students of the ChatGPT group were encouraged to experiment with it; ask different questions; ask for videos, images, and internet resources; and in general to be creative, imaginative, and playful while using this new tool. Once they had the final AI content, they were advised to critically evaluate it by comparing it with the relevant content of a reliable scientific resource, such as a textbook or published article, and perform the necessary modifications to the AI output. After finishing the assignment, they were asked to complete an open-ended questionnaire individually (Multimedia Appendix 1), including questions about the usability, problems, opinions, proposals, and so forth, which was emailed to them, and which they would submit to the educator together with the assignment (ie, the PowerPoint presentation).

The AI Evaluation Questionnaire included 12 questions requiring free-text responses and was developed by the authors by combining questions from 2 sources: essays evaluation questionnaires retrieved in the scientific literature [22-24] and the questionnaire ChatGPT produced on the prompt “Can you develop 10 questions for a user to evaluate your performance on writing an essay?” Questions were combined and modified, they were piloted within a small student group other than the research groups, and they were finally amended as necessary. The free-text comments of the AI Evaluation Questionnaire were grouped into main themes and discussed (subjective and qualitative evaluation).

After students completed and submitted their projects via email, and on the designated day they would present the PowerPoint presentations in class, at the beginning of the session, they all had an unannounced blind knowledge examination (answered individually and anonymously, where they only indicated the group they belonged in, so that the educator could not relate the students with the answer sheets). The examination was developed by the authors and consisted of 10 multiple-choice questions (MCQs), which addressed the learning objectives of the topic. They were informed that the knowledge test was intended for the educator to identify whether the assignment had equipped them with the intended knowledge and whether there were any knowledge gaps to address. The results of the
examination (examination grades) were compared among the 2 groups, that is, the literature research group and the ChatGPT group. Statistically significant differences between the groups’ grades were explored using the Mann-Whitney nonparametric test. Data analysis was conducted using SPSS (version 25.0; SPSS Inc), and statistical significance was set at $P=.05$ (objective and quantitative evaluation).

The final study design is summarized as follows:

- Students were randomly divided into 2 large groups (the ChatGPT and the literature research groups) and further into smaller groups.
- Literature research group performed the assignment by searching the internet and delivered it in PowerPoint format, including the references used.
- ChatGPT group (1) asked the LLM relevant queries and developed a PowerPoint presentation; (2) registered and reported on their interactions with ChatGPT, including the prompts and their modifications, the final outcome and its evaluation after comparing it with a reference text or book chapter; and (3) answered the AI Evaluation Questionnaire on their experience with the LLM.
- All students presented their learning assignments in class. At the beginning of this session, they undertook an unannounced knowledge examination of 10 questions.
- Data derived from the knowledge examination grades, the PowerPoint presentations, and the free-text comments of the AI Evaluation Questionnaire.

### Results

#### Quantitative Results

Out of the 77 students, 39 were assigned to the ChatGPT group forming 9 subgroups and 38 to the literature research group forming 8 subgroups. Seventy students undertook the MCQ examination (7 students were absent) and examination grades ranged from 5 to 10 on the 0-10 grading scale. Figure 1 presents the number of students (percentages within each group) with their examination grades. We noticed that in the higher range of examination grades, that is, 8-10, the ChatGPT students outperformed the literature research students, while the opposite happened within the lower range of examination grades, that is, 5-7.

To check for differences between the ChatGPT student group and the literature research group, we performed the Mann-Whitney $U$ test, which showed that students of the ChatGPT group ($n=39$; mean 7.54, SD 1.18) performed significantly better ($P=.045$) than students in the literature research group ($n=31$; mean 6.94, SD 1.12).

To foster inclusiveness and avoid discrimination, we deliberately chose not to perform statistical analyses regarding gender differences, as we also believe that gender diversity is not associated with the educational process or the educational outcomes. Education is offered equally to all students and any gender differences possibly found would not differentiate educational approaches for one gender or the other. Instead, we perceive this student cohort as representatives of their generation (Generation Z), a characteristic that is directly related to this study’s outcomes and could explain several findings. This concept is in line with the US National Institute of Health recommendations for gender-neutral language [25].

![Figure 1. Students’ examination grades (% of students within each group).](https://mededu.jmir.org/2024/1/e51344)
Qualitative Results

Overview
Out of the 39 students of the ChatGPT group, 31 (80%) students answered the 12 questions of the AI Evaluation Questionnaire. The free-text answers to the questions were grouped into themes and discussed. Three main themes emerged.

Collaboration With ChatGPT and Problems Encountered
Although the majority of students were aware that ChatGPT had surfaced a couple of months ago in the digital world and some of them had already used it, this was the first opportunity they had to actually work with it and “officially” use it within their studies, and they enjoyed and appreciated this opportunity. They characterized it as a “powerful and versatile tool,” “intuitive and intelligent,” “revolutionary,” and “enjoyable to work with” and they thought this experience was “interesting and different from the regular assignments.” They stated that learning to use these AI tools would improve their future practice but emphasized that “you have to learn how to properly use it.” They appreciated its human-like answers, as these “do not make the user feel distanced from technology.” A student stated:

In the beginning I was afraid it was going to be too difficult to work with but as I was discussing with it I understood its greatness. I think it really is the future as it can help both education and research. I really did enjoy its human-like answers like when something was wrong it persisted like a human being for its accuracy as well as when it did not answer the question as it should like a lazy student.

Another student commented: “I enjoyed working with ChatGPT, because I got to learn and understand something that is going to be a part of the future.” Humanization of the LLM is worth noting: “He always understood what we wanted.” Textbox 1 shows examples of students’ prompts.

Textbox 1. Examples of students’ prompts to ChatGPT (exact copies).

- How does radiation affect human health?
- What’s the difference between deterministic & stochastic effects of radiation?
- Is radiation exposure carcinogenic?
- Which are the radiation doses from common dental radiographic exams?
- Which criteria are used to reduce unnecessary radiographic exposure in dentistry?
- Can a pregnant employee continue to work in the dental radiology department?
- What is the importance of radiation biology? With references used
- What are the effects of radiation on cells and tissues? With references used
- What are the effects of radiation on the oral cavity? Rewrite the previous answer in a more elaborate way
- Make a chart about effective dose from diagnostic x-ray examinations focusing on the oral cavity
- Radiation biology, include references
- Measurements of radiology safety, include references
- Radiology protection in dentistry, include references
- How can we minimize the radiation exposure on dental staff, including references
- Why are radiation safety precautions necessary for the dentist
- Tell me how radiation can affect the human body
- Write me an essay discussing radiology safety and protection procedures in dentistry
- Can you explain radiation biology for medicine and dentistry in 400 words, include references
- Radiation exposure in dental office word limit 200-250 words. Include references
- Radiation monitoring in the dental office in 230-270 words include references
- Write me an essay of 400 words about the biology of radiation and provide references
- Write me a 300 words essay about radiation safety and protection in dentistry
- What are the risks associated with exposure to radiation?
- What are the modifying factors of irradiation?
- How does radiation exposure time and dose differentiate between adults and children in dental x-ray taking?

Not unexpectedly, students identified all the problems and limitations of ChatGPT, which are later described in detail in the literature. They identified the need to rephrase or detail the prompts to have a satisfactory output (“we learned quickly how to ask the questions to get a good answer”) and realized that if the same question was asked slightly differently the output was
different ("by asking it 6 different questions, we wanted to get a better idea of what it changes on the text every time we put a new word or phrase the question differently"). They confirmed that some information was outdated, important content was missing, part of the answer was occasionally incorrect, links to references were nonexistent, and the links to videos were not working, although the LLM provided detailed and seemingly reliable information on the links and references (thus unknowingly identifying the "hallucination" effect of ChatGPT).

A student stated: “Mostly it understood our questions but it was not giving us that detailed and satisfactory answers as we anticipated according to our book.” Another student correctly noticed that “ChatGPT is not capable of having thoughts or opinions on its own, so it does not answer some questions that demand a critical-thinking answer.” Technical issues were also mentioned by some students, for example, “some days it was not opening and our conversation couldn’t be saved on the cloud” and “it ‘crushed’ sometimes mid-working.”

**Quality of the Generated Outputs**

Students found that the quality and depth of the information provided by ChatGPT depended on the quality and wording of the questions asked. As a student noticed:

> I would not say that it demonstrated a very deep understanding of the topic, but I think with even more questions being asked, the text could essentially show a deep understanding of the topic.

Students quickly realized that with follow-up questions and rewording, they could guide the LLM to produce more detailed and in-depth answers: “it needed some guidance with follow up questions to further specify what we were asking for.” While comparing the output with a reference text, students reported that the answers were not detailed; sometimes included false data; and were brief, general, or superficial; nevertheless, the key points were evident. A student concluded that “ChatGPT is more than enough in order to understand and have a general idea about the main points of the matter being discussed” and another student thought that “I will find more details by going and searching online or in books.” They expect ChatGPT to improve in the future and be able to provide videos and images because “they are helpful in understanding a topic and provide a more effective way to retain information as well” and also to be able to browse external resources outside its stable database (Figure 2).

They evaluated the language as appropriate for a scientific document, understandable, and explanatory, and they indicated that when references were asked for, the language was even more formal and academic: “It is fascinating how the AI provides understandable answers in a scientific manner.” However, they encountered problems with the references, as in some occasions, ChatGPT denied to supply them, while in other instances, the references were incorrect. A student described:

> The AI was continuously denying to give us relative references but after reforming our questions we eventually got our answer. The references it used were accurate scientific resources found on its stable database like the American Dental Association.

Another student stated that “We used chat GPT 4 so all our references were sufficient and up to date” (apparently overestimating ChatGPT-4’s currentness, as it has the same cutoff date as ChatGPT-3.5). The majority of students evaluated the references as relevant, sufficient, reliable, and up-to-date; however, they also recognized the limitations of the LLM, thinking that “it is under construction so not all its answers are up to date and sufficient information is only provided up to a certain point in time.”
Exploring Additional Possibilities and Predicting the Future

Students experimented with ChatGPT, asking it to provide images and videos, and create MCQs, charts, bullet point summaries, and presentation templates, for example, “we asked about multiple choice questions and the answers were actually impressive” (Figure 3). Students were imaginative and resourceful, and they were disappointed when their request was not realized:

I asked from it to provide me some explanatory images related to our topic, but it was not able to do so. I think this is a crucial disadvantage, as images give depth and context to a description and provide a much more immersive experience than writing alone.

Two student groups—comprised of technologically very experienced students—surprised the authors when they skillfully bypassed the inability of ChatGPT to produce PowerPoint presentations by asking it to write a programming code:

We used the AI for the generation of a PowerPoint. Since it cannot on its own generate PowerPoint Slides we asked it to generate a VBA code for the PowerPoint. That code was copied and then pasted to the ‘Developer’ section of the PowerPoint. As a result we got a beautiful but not so detailed presentation of our topic.

This process enabled the instant transfer of ChatGPT’s output within a PowerPoint slide presentation created by ChatGPT. Among the future applications of ChatGPT, students included the use in dental education, for example, for the creation of MCQs, summarizing a topic, lecture revision, helping students better understand a theory or concept, assignments and projects, laboratory reports, questions about law and ethics, communication with patients, and more. A student proposed:
Virtual patient consultations: ChatGPT could be used to simulate patient consultations for dental students. Students could practice various scenarios, including patient history taking, explaining diagnoses, and treatment planning.

Continuing education could also avail from the opportunities ChatGPT and LLMs offer:

Education that never ends: ChatGPT may be utilized to give dental professionals continual education. For dental professionals to keep current in their field, faculty might create modules containing the material they need, and ChatGPT may offer engaging tasks and tests to reinforce the learning.

Considering dental practice, students proposed that ChatGPT could be used to educate and solve problems for the dentist, for example, when “the dentist has a mind block” or when the dentist “seeks information about new dental materials and techniques”; also for treatment plans, schedule creation, and oral hygiene info; and for patient education “through integrating the model into a dental practice’s website or patient portal.”

For research and scientific publications, students thought it “can be useful to use it synergistically with your own research,” but “you should always double-check the information” and “keep in mind the plagiarism, using the information provided appropriately.”

Finally, students admitted that ChatGPT has drawbacks such as a limited database, incapability to access external web resources and provide images and videos, inaccurate links, and the need to verify the information generated. They thought that “it should be used with caution” and that “AI still needs to evolve,” so that it will become “an incredibly smart, effective, and powerful tool that can help the scientific community.” They realized that “the power it holds is unpredictable and the work of doctors could be compromised” and feared that “maybe we will live one day that AI robots could even replace dentists.” A student eloquently summarized ChatGPT’s past, present, and future:

After many years of research and after many science fiction movies about the power of AI and its impact on society I have come to the conclusion that this kind of AI can only help and do no harm. AI like ChatGPT that is available to the public and gives sufficient and accurate responses can give us hundreds of possibilities, even at dentistry. But I really don’t know this exact ChatGPT with its limited dental references can influence the field of dentistry. I can though imagine a more resourceful AI where it uses PubMed or Research Gate to generate its responses that would really elevate the level of dental education and research. What if a curious dentist had the million dollar question answered in milliseconds by the AI?

Figure 3. Multiple-choice questions created by ChatGPT. MCQ: multiple-choice question.
Discussion

Overview

In March 2023, a total of 39 dental students who are 20 years of age, through composing an educational assignment, identified the capabilities and limitations of the recently introduced ChatGPT and explored various possibilities; used it to write MCQs and programming codes; proposed future applications in education, research, and dental practice; and outperformed their peers in the knowledge examination.

Results Explained and Compared

The quantitative results, that is, the examination grades, demonstrated that all students performed well (their grades fell within the middle and high ranges of the grading scale) and no students underperformed (no grades in the low ranges of the scale), while ChatGPT group students outperformed their literature research group peers. Since the examination occurred with no prior notice to the students, it directly reflects the knowledge acquired and retained through the project’s creation. Students’ good performances on the examination could be related to the format of the project in connection with their generational traits: all students socially belong to the Generation Z cohort (born between 1995 and 2010), so they are the first true “digital natives” [26], having grown up with smartphones, social networks, apps, and streaming content as part of the daily routine [27]. They are considered tech-savvy, mobile-driven, collaborative, and pragmatic [28,29] and possess a natural facility with digital tools and an interest for everything digital. Motivated by the opportunity to use the internet and work collaboratively, students immersed themselves in the project and explored it in depth, and this applies even more to the ChatGPT group students who were excited and curious to test this new digital tool. The enhanced learning observed with the ChatGPT students can be also attributed to the increased “time on task” for these students, as they had to spend more time asking and reasking the questions, evaluating the answers, correcting, and complementing them in comparison to their peers who had clear and readily available results from the relevant scientific literature. Additionally, ChatGPT group students had to work more than their fellow students with the learning material at a higher cognitive level and constantly apply critical thinking while experimenting with various questions and answers, comparing, and synthesizing them—an element that also enhances deep learning and results in enhanced performance [30].

The AI Evaluation Questionnaire provided insight into students’ opinions, evaluations of ChatGPT, the problems encountered, and their future estimations. Students demonstrated their prescience by providing remarks in concordance with those found in later-published articles; the latter were accessed by the authors after the research was concluded and while composing this study. Students evaluated their learning experience with ChatGPT as interesting, enjoyable, and engaging [19] and appreciated its user-friendly interface and the possibility of arguing with it [4,16]. They assessed the generated content as overall correct and sufficient [7,31], although often providing a general overview of the subject [5], as well as not demonstrating a deep understanding of the context [32-34] nor thinking critically [10,35]. They first-hand identified the need for carefully created questions [36] and critical analysis of the answers [14,36], and they urged for cautious and responsible use [4,6]. In agreement with Chan and Hu [11], they are ready to embrace this new technology but in a collaboration where people maintain control and are not replaced by AI [17,20,37,38]. Finally, in line with the literature, they attributed “anthropomorphic” qualities to the language model (1 student referred to ChatGPT using the gender pronoun “he”), possibly explained by the establishment of a personal connection between the student and the language model while engaging in human-like conversations in combination with student’s own gender-related perceptions and interaction style [39].

Students proposed possible applications of ChatGPT in education for revisions, MCQ creation, personalized learning, writing essays [3,4,20,37,40], and continuing education [38], as well as in research and clinical practice [4,6,12]. Nevertheless, students thought that the LLM must evolve to provide images, videos, accurate and relevant citations, and browse the internet [31,41,42].

Numerous publications thereafter examined the LLM’s limitations that had been already identified by the students: incorrect answers and outdated content [10] possibly due to its limited data set [37,38,43], the possibility for fabricated information and hallucination [44], false citations and links leading to nonexistent sources [38,44,45], inability to browse the web [41], and risks for plagiarism [3,46].

This research materialized Kung et al’s [31] concluding remarks that “the utility of generative language AI for medical education must be studied in real-world learning scenarios with students, across the engagement and knowledge spectrum” since ChatGPT was embedded within the educational process, thus producing authentic and relevant results. The quantitative and qualitative outcomes of this study indicate that this cohort of Generation Z students is capable of adapting quickly to new technologies and ready to use LLMs such as ChatGPT in the learning process—while acknowledging their limitations—particularly when these tools are integrated within a pedagogical framework that fosters creativity and autonomous learning. Educators on the other hand seem to have limited technological knowledge, skills, and pedagogical expertise to assess AI applications and successfully integrate them into education [12,47]; therefore, they should pursue professional development to develop new skills related to AI understanding, possibilities, and implementation [15,40,48,49].

Pedagogical Aspects

All second-year students were asked to explore the topic of “Radiation Biology and Radiation Protection in the Dental Office” and develop assignments to be presented in class as PowerPoint presentations. Questions and knowledge gaps were covered during the in-class presentations by the instructor and not infrequently by their peers. This approach is consistent with the “flipped classroom” concept, an educational methodology that research has shown to engage students in the learning process, promote autonomy and self-regulation, allow for higher-order thinking, improve student satisfaction, and increase
academic performance [50,51]. Another element of pedagogical interest is the small group collaborative work to develop the assignments. Collaborative learning has the potential to promote deep learning, which is essential for understanding complex concepts particularly in science education, through students’ meaningful interactions and constructive debates [52]. Scager et al [52] reported that effective collaboration is achieved when students undertake a challenging, complex task, and they succeed in creating a new and original output. Such tasks applied in higher education build a sense of responsibility and shared ownership of the output and the collaborative process, and this sense was indeed apparent in the students of this study within and during their oral presentations.

An additional pedagogical element is the learning assignments as a method for self-learning and knowledge acquisition. Learning through assignments has been reported to be preferred by students: in the study of Warren-Forward and Kalthoff [53], 79% of the students reported that the assignment on magnetic resonance imaging safety was both a positive learning experience and provided an understanding of the topic. Writing assignments enhance retention of knowledge; when assignments include reflective thinking, for example, when students have to evaluate and synthesize information (as happens in this study), higher-order (critical) thinking is also enhanced as students work at a higher cognitive level [30].

The innovative pedagogical aspects of this study (flipped classroom, learning assignments, and group learning) constituted a supportive environment for students of both groups to demonstrate their skills, achieve the learning objectives, and produce valuable results. While this pedagogical approach may cater more to certain types of learners, it remains pertinent for younger generations, who prefer active and collaborative learning.

**Study Design: Tackling the Challenges**

Of interest would be to communicate herein the challenges faced during designing the research process, as the ChatGPT environment was largely unknown at the time, and obstacles and drawbacks had to be identified and resolved ahead through a step-by-step prospective analysis of the sequence of events.

For example, a concern that had to be addressed ahead was the fact that the subject was unknown to the students and they would not know whether the output was scientifically correct or incorrect, comprehensive or incomplete because they would not have an exemplary scientific text to compare it with, as they would rely solely upon ChatGPT’s answers. To address this, they were advised to compare the outcome with the relevant content of a recommended textbook (or other reliable source of their choice), critically evaluate the quality of the AI outcome, and perform the necessary amendments to complement or correct the AI results. The comparison should be included either within their presentation or within the AI Evaluation Questionnaire. This process would additionally ensure the achievement of learning objectives. In line with this process and at a later time, Chung [48] proposed in his article published in April 2023 that “instructors should teach students to use other authoritative sources (e.g., reference books) to verify, evaluate, and corroborate the factual correctness of information provided by ChatGPT.”

Another concern arose about elucidating students’ engagement with ChatGPT: since the output of ChatGPT would be texts in slide format (similar to the ones of the literature research group), the educator (one of the authors) could evaluate these texts or slides for accuracy and comprehensiveness but could not comprehend whether they were generated following single or multiple attempts, posing differentiated or follow-up queries; therefore, the time and effort spent on the research process and the learning path could not have been assessed nor would the capabilities and drawbacks of the LLM be revealed. To address this concern, the ChatGPT group students were asked to register and report all their interactions with the LLM (including the number of prompts, the modification of prompts, the queries about references, images, and the underlying reasoning); thus, the educator could evaluate the cognitive effort they put in the assignment and the critical thinking applied until a satisfactory result was achieved. Furthermore, this would provide valuable insights into comprehending the usability and operational characteristics of the LLM. Adding to this, the AI Evaluation Questionnaire was a useful means to draw information on student-LLM interactions.

In accordance with the above procedure determined by the authors and in affirming their decisions, Halaweh’s study [14] published in April 2023—2 months after the development of this study’s design and 1 month after its implementation—precisely described the same process when discussing the strategies for successful implementation of ChatGPT in education. It seems that future literature confirmed the authors’ study design overall.

**LLMs in Higher Education**

Given the study’s results and in agreement with the relevant literature, the authors would suggest that higher education institutions and dental schools could consider updating their curricula, policies, and teaching methods to prepare students for an AI-driven future, by including education on and with AI tools and LLMs [8,45]. Within this context, faculty professional development seems urgent to increase their skill level and AI understanding, for example, through peer support, mentoring, and sharing good teaching practices [36], as most educators have limited knowledge and skills to assess and efficiently use AI applications [12]. The introduction of LLMs into education will offer opportunities to improve its efficiency and quality: improved student performance, personalized learning, targeted and immediate feedback, increased accessibility, creativity and innovations, student engagement, lesson preparation, collaborative activities, and evaluation [4,40,54-56]. From the pedagogical perspective, students using LLMs have the potential to develop new competencies including 21st-century soft skills, such as self-reflection abilities, problem-solving skills, creative and critical thinking, and collaboration, thus becoming motivated and autonomous learners [3,4,16,33,49]. Moreover, as AI technology evolves and gradually integrates within the educational process, the conventional pedagogical theories may not be relevant nor sufficient to support the teacher-student-technology relationship, as technology
profoundly alters the way students learn and engage with the content and the teacher; innovative pedagogies will be needed, such as the “entangled pedagogy” Fawns [57] proposed to contextualize students’ learning in a world where AI is increasingly prevalent [15,16].

To respond to the AI paradigm shift, higher education institutions, educators, and students must engage in constructive dialogue to develop policies, guidelines, and training opportunities for the implementation of innovative technological tools in the teaching process [16,34,55]. Despite the current weaknesses that limit their implementation, LLMs will likely improve in the future in terms of performance, scalability, and quality of responses, as well as through fine-tuning for specific tasks, customized use cases, and search engine connection [4,16,31,58].

Limitations and Strengths
The small number of students who participated in this study (77 in total and 39 in the ChatGPT group) in 1 dental school can limit the extrapolation of the results. Students’ digital literacy is also of relevance: students who participated in this research were mostly tech-savvy, whereas students in other schools or universities may be less familiar with digital technologies; thus, results would not apply to them [17]. In addition, some findings (particularly the qualitative ones) may be outdated at the time of publication, as LLMs constantly evolve and new LLMs have been introduced since the research was conceptualized and implemented. For example, Google Bard and Microsoft Bing claim to have live access to the internet, a capability highly appreciated by the students; ChatGPT has since evolved its algorithms, with results being more accurate and relevant. Some elements of the study design could have been further explored; for example, students’ assignments could have been graded and compared, but since assignments’ grading was not included in the semester program of the module, this was not performed. In any case, the importance of this study lies in the fact that this was a very early attempt to implement legitimately and in vivo a language model in the teaching process as a partner in learning, in contrast to the large number of publications perceiving ChatGPT as a partner in cheating and academic dishonesty [12,59,60]. Another strength would be that it revealed aspects of the language model-students’ interactions during the learning process, which indicate that this emerging relationship is yet to be explored, and updated pedagogical frameworks are needed for this purpose.

Conclusions
ChatGPT was implemented in real-life undergraduate dental education and was evaluated. Students using ChatGPT for their learning assignments performed significantly better in the knowledge examination than their fellow students who used the literature research methodology. The AI questionnaire answered by students revealed the capabilities and weaknesses of the language model, as identified later in the scientific literature. Students enjoyed working with this tool and explored different options and possibilities, indicating that they are technologically knowledgeable and capable of adapting to new technologies, both in education and in future clinical practice. LLMs such as ChatGPT have the potential to play a role in education, underpinned by solid pedagogies.

Acknowledgments
The authors are grateful to the students who participated in the study. They were enthusiastic, motivated, and resourceful and explored the subject in depth, thus providing valuable insights to inform the ongoing research on the topic.

Authors’ Contributions
AK conceptualized, designed, and realized the study; interpreted the data; and drafted the manuscript. KG supervised the project, reviewed the literature, and contributed to drafting the manuscript. MADdS and EGK critically reviewed and revised the manuscript; EGK performed the statistical analysis. VS consulted on information technology and reviewed the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
AI evaluation questionnaire.
[DOCX File, 14 KB - mededu_v10i1e51344_app1.docx ]

References


22. IL rubric for student essay evaluation. wordpress.com. URL: https://smcmaiprject.wordpress.com/methods/rubric/ [accessed 2023-12-21]

23. FREE 39+ student evaluation forms in PDF | Excel | MS word. SampleForms. URL: https://www.sampleforms.com/student-evaluation-form-template.html [accessed 2023-12-21]


27. 7 unique characteristics of generation Z. Oxford Royale. URL: https://www.oxford-royale.com/articles/7-unique-characteristics-generation-z/ [accessed 2023-12-21]


**Abbreviations**

- **AI**: artificial intelligence
- **LLM**: large language model
- **MCQ**: multiple-choice question
Increasing Realism and Variety of Virtual Patient Dialogues for Prenatal Counseling Education Through a Novel Application of ChatGPT: Exploratory Observational Study

Megan Gray1, MD; Austin Baird2, PhD; Taylor Sawyer1, MBA, MEd, DO; Jasmine James3, MPH, MD; Thea DeBroux1; Michelle Bartlett4, MS, MD; Jeanne Krick5, MA, MD; Rachel Umoren1, MS, MBBCh

1Division of Neonatology, University of Washington, Seattle, WA, United States
2Division of Healthcare Simulation Sciences, Department of Surgery, University of Washington, Seattle, WA, United States
3Department of Family Medicine, Providence St Peter, Olympia, WA, United States
4Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, United States
5Department of Pediatrics, San Antonio Uniformed Services Health Education Consortium, San Antonio, TX, United States

Corresponding Author:
Megan Gray, MD
Division of Neonatology
University of Washington
M/S FA.2.113
4800 Sand Point Way
Seattle, WA, 98105
United States
Phone: 1 206 919 5476
Email: graym1@uw.edu

Abstract

Background: Using virtual patients, facilitated by natural language processing, provides a valuable educational experience for learners. Generating a large, varied sample of realistic and appropriate responses for virtual patients is challenging. Artificial intelligence (AI) programs can be a viable source for these responses, but their utility for this purpose has not been explored.

Objective: In this study, we explored the effectiveness of generative AI (ChatGPT) in developing realistic virtual standardized patient dialogues to teach prenatal counseling skills.

Methods: ChatGPT was prompted to generate a list of common areas of concern and questions that families expecting preterm delivery at 24 weeks gestation might ask during prenatal counseling. ChatGPT was then prompted to generate 2 role-plays with dialogues between a parent expecting a potential preterm delivery at 24 weeks and their counseling physician using each of the example questions. The prompt was repeated for 2 unique role-plays: one parent was characterized as anxious and the other as having low trust in the medical system. Role-play scripts were exported verbatim and independently reviewed by 2 neonatologists with experience in prenatal counseling, using a scale of 1-5 on realism, appropriateness, and utility for virtual standardized patient responses.

Results: ChatGPT generated 7 areas of concern, with 35 example questions used to generate role-plays. The 35 role-play transcripts generated 176 unique parent responses (median 5, IQR 4-6, per role-play) with 268 unique sentences. Expert review identified 117 (65%) of the 176 responses as indicating an emotion, either directly or indirectly. Approximately half (98/176, 56%) of the responses had 2 or more sentences, and half (88/176, 50%) included at least 1 question. More than half (104/176, 58%) of the responses from role-played parent characters described a feeling, such as being scared, worried, or concerned. The role-plays of parents with low trust in the medical system generated many unique sentences (n=50). Most of the sentences in the responses were found to be reasonably realistic (214/268, 80%), appropriate for variable prenatal counseling conversation paths (233/268, 87%), and usable without more than a minimal modification in a virtual patient program (169/268, 63%).

Conclusions: Generative AI programs, such as ChatGPT, may provide a viable source of training materials to expand virtual patient programs, with careful attention to the concerns and questions of patients and families. Given the potential for unrealistic or inappropriate statements and questions, an expert should review AI chat outputs before deploying them in an educational program.
In this report, we explore the use of ChatGPT to enhance the realism of the VANESSA VSP. We hypothesized that the integration of the ChatGPT AI chatbot would generate realistic, relevant, and usable patient responses for a VSP simulator used in prenatal counseling education.

Methods

The study used an exploratory observational design, with ChatGPT acting as an expectant parent within the VANESSA software, conducted in February 2023 on ChatGPT 3.5.

The VANESSA VSP represents a pregnant woman in her 23rd week of gestation and showcases emotions through animated facial expressions and body language. Created with input from neonatologists, its dialogue and emotive feedback were found realistic in pilot tests, enabling participants to confidently identify its emotional states.

ChatGPT is a large language model developed by OpenAI. Its exceptional performance stems from generative pretraining, leveraging extensive unlabeled data sets [10]. This foundational training helps it grasp English nuances. Following this pretraining is “one-shot” learning, a rapid task-specific learning [11]. The architecture includes a transformer encoder-decoder neural network, originally developed for translation services and now popular in language models [12]. ChatGPT decodes user prompts to create relevant responses using autoregressive language modeling [13]. It is apt for generating realistic dialogue for health care simulations.

The study had three phases:

1. ChatGPT generated a list of common concerns from families expecting preterm delivery at 24 weeks. The stability of these concerns was verified in an iterative process over time.
2. Using a standardized prompt, ChatGPT crafted potential parent questions related to each concern.
3. Role-plays were constructed for a mother expecting preterm delivery. The AI was given varied emotional settings for the scenario of preterm labor at 24 weeks, including anxiety or distrust in the medical system. In each scenario, a designated area of concern and a primary question were specified, derived from the potential parent responses generated in phases 1 and 2. Conversations were created with cues for the VSP and then reviewed (an example is shown in Figure 1).
Throughout the process, ChatGPT was instructed to adhere to a fifth-grade reading level for the AI parent role. Considering the US Department of Education’s findings on widespread low literacy, the importance of health literacy, and the impacts of pain, stress, and other factors on comprehension, this was deemed crucial [14-17]. The Joint Commission and several medical organizations suggest that patient materials should be at a fifth-grade level or lower [18]. Although these dialogues were verbal, the principle of understandability remained in place.

Role-play conversations were scrutinized for parental responses. Each was checked for question or statement content, emotional cues, and sentence count. Initially, generated physician names and certain response starters (eg, “yes” or “no”) were noted but removed for evaluation. Sentences were then appraised by a neonatologist for realism, relevance, and usability for virtual prenatal counseling simulations. Each metric used a 5-point Likert scale, ranging from 1 (the lowest) to 5 (the highest). For usability in the VANESSA VSP, responses were scored as follows: 1 if they were unusable, 2 if they were unusable without major modifications, 3 if they were usable with moderate modifications, 4 if only minor modifications were needed, and 5 if they were usable without any modifications. The first 10% of responses were independently reviewed by 2 experienced neonatologists (RU and MG) and then compared for reliability. A calculated weighted kappa on the sample was 0.84, which is considered a strong level of agreement [19]. Responses with differences in rating were discussed by the team members to improve reliability, and the remainder of the data set was scored by one of the experienced neonatologists. Duplicate responses were scored only once. Analysis was done using Stata (version 17.0; StataCorp).

**Results**

ChatGPT-3.5 generated a list of 7 common areas of concern, 28 questions likely to be asked by parents anxious about the preterm delivery of their infant, and 7 additional questions likely to be asked by parents with low trust in the medical system (Table 1). These areas of concern and questions were used to create 35 unique role-plays, which contained 176 unique parent responses (Table 2). The role-plays had a median of 5 (IQR 4-6) parent responses to the counseling physician. The responses were roughly evenly split between questions and statements. About half of the responses had 2 or more sentences in the response. Many responses mentioned a specific emotion or feeling. The role-play of the parent with low trust in the medical system generated 50 unique sentences across the 7 areas of concern. There was variation in the number of unique sentences generated across the 7 major areas of concern (Table 3). Most responses were found to be realistic, appropriate for variable conversation paths, and usable in a VSP program (Table 4).
Table 1. Areas of concern and example questions generated by artificial intelligence.

<table>
<thead>
<tr>
<th>Areas of concern</th>
<th>Example questions from parents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health and development</td>
<td>• Will our baby be healthy if they are born too soon?</td>
</tr>
<tr>
<td></td>
<td>• What will the doctors do to help our baby be healthy and strong?</td>
</tr>
<tr>
<td></td>
<td>• Can our baby get sick more easily if they are born too soon?</td>
</tr>
<tr>
<td></td>
<td>• Will the baby feel pain during birth or while in the hospital?</td>
</tr>
<tr>
<td></td>
<td>• I’m worried about the risks and complications, what if something goes wrong? (Mistrust)</td>
</tr>
<tr>
<td>Survival</td>
<td>• Will the baby survive?</td>
</tr>
<tr>
<td></td>
<td>• What kind of help will our baby need to stay alive?</td>
</tr>
<tr>
<td></td>
<td>• How likely is it that our baby will survive?</td>
</tr>
<tr>
<td></td>
<td>• What kind of machines or medicines will our baby need to help them breathe and stay alive?</td>
</tr>
<tr>
<td></td>
<td>• I don’t know if I can trust the medical field, what are the chances of my baby surviving at 24 weeks? (Mistrust)</td>
</tr>
<tr>
<td>NICU stay</td>
<td>• What is the NICU, and why does our baby need to go there?</td>
</tr>
<tr>
<td></td>
<td>• How long will our baby need to stay in the NICU?</td>
</tr>
<tr>
<td></td>
<td>• Can we visit our baby in the NICU, and how often?</td>
</tr>
<tr>
<td></td>
<td>• Will our baby be alone in the NICU, or will there be other babies and parents there too?</td>
</tr>
<tr>
<td></td>
<td>• What kind of things can we do to help our baby feel better in the NICU?</td>
</tr>
<tr>
<td></td>
<td>• Will anything happen in the NICU without my consent? (Mistrust)</td>
</tr>
<tr>
<td>Emotional impact</td>
<td>• How do we get ready for having a baby born too soon?</td>
</tr>
<tr>
<td></td>
<td>• Can we hold and touch the baby in the hospital, and is this good for the baby?</td>
</tr>
<tr>
<td></td>
<td>• Who can help us if we are feeling sad or stressed about our baby being born too soon?</td>
</tr>
<tr>
<td></td>
<td>• I’m worried about my baby going to the NICU where she will be alone and scared (mistrust).</td>
</tr>
<tr>
<td>Long-term outcomes</td>
<td>• What help can we get after we leave the hospital?</td>
</tr>
<tr>
<td></td>
<td>• Will our baby be able to do the same things as other babies who were born at the right time?</td>
</tr>
<tr>
<td></td>
<td>• Will our baby be okay in the future if they are born too soon?</td>
</tr>
<tr>
<td></td>
<td>• I don’t know what’s going to happen to my baby. I don’t really trust the doctors but what happens if my baby doesn’t develop properly? (Mistrust)</td>
</tr>
<tr>
<td>Feeding and nutrition</td>
<td>• How will our baby get the right kind of food if they are born too soon?</td>
</tr>
<tr>
<td></td>
<td>• Can we feed our baby ourselves, or will they need special milk or formula?</td>
</tr>
<tr>
<td></td>
<td>• How often will our baby need to be fed, and how much?</td>
</tr>
<tr>
<td></td>
<td>• Will our baby be able to eat the same kinds of food as other babies when they get older?</td>
</tr>
<tr>
<td></td>
<td>• Can we breastfeed our preterm baby, or do we need to use formula?</td>
</tr>
<tr>
<td></td>
<td>• Will our baby be able to breastfeed right away, or will they need to be fed in a different way at first? (Mistrust)</td>
</tr>
<tr>
<td></td>
<td>• Will I have any say in how my baby is fed? (Mistrust)</td>
</tr>
<tr>
<td>Quality of life</td>
<td>• Will our baby be able to go to school and play sports like other kids?</td>
</tr>
<tr>
<td></td>
<td>• How can we help our baby if they have trouble learning or doing things in the future?</td>
</tr>
<tr>
<td></td>
<td>• What can we do to make sure our baby has the best chance for a good future?</td>
</tr>
<tr>
<td></td>
<td>• I’ve had bad experiences before and I’m scared about what’s going to happen to my baby in the future, what can I expect? (Mistrust)</td>
</tr>
</tbody>
</table>

aNICU: neonatal intensive care unit.
### Table 2. Generated role-plays by artificial intelligence.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Role-plays (n=35), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Worried about specific area of concern</td>
<td>28 (80)</td>
</tr>
<tr>
<td>Low trust in the medical system</td>
<td>7 (20)</td>
</tr>
<tr>
<td>Responses per role-play, median (IQR)</td>
<td>5 (4-6)</td>
</tr>
<tr>
<td><strong>Parent responses (n=179), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Unique responses</td>
<td>176 (98)</td>
</tr>
<tr>
<td>Duplicate responses</td>
<td>3 (1)</td>
</tr>
<tr>
<td><strong>Types of responses (n=179), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Statements</td>
<td>91 (51)</td>
</tr>
<tr>
<td>Questions</td>
<td>88 (49)</td>
</tr>
<tr>
<td><strong>Sentences per response (n=179), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>81 (45)</td>
</tr>
<tr>
<td>2</td>
<td>76 (42)</td>
</tr>
<tr>
<td>3</td>
<td>18 (10)</td>
</tr>
<tr>
<td>4</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Duplicate sentences (n=305), n (%)</td>
<td>37 (12)</td>
</tr>
<tr>
<td>Total unique sentences (n=305), n (%)</td>
<td>268 (88)</td>
</tr>
<tr>
<td><strong>Feelings stated in responses (n=117), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Specific emotion stated in phrase</td>
<td>56 (48)</td>
</tr>
<tr>
<td>“Scared”</td>
<td>36 (31)</td>
</tr>
<tr>
<td>“Worried”</td>
<td>26 (22)</td>
</tr>
<tr>
<td>“Anxious”</td>
<td>2 (2)</td>
</tr>
<tr>
<td>“Concerned”</td>
<td>2 (2)</td>
</tr>
<tr>
<td>“Afraid”</td>
<td>1 (1)</td>
</tr>
<tr>
<td>“Nervous”</td>
<td>1 (1)</td>
</tr>
<tr>
<td>“Overwhelmed”</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Emotion indirectly implied by phrase</td>
<td>51 (44)</td>
</tr>
</tbody>
</table>

### Table 3. Sentences generated per role-play.

<table>
<thead>
<tr>
<th>Area of concern</th>
<th>Number of unique sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health and development</td>
<td>47</td>
</tr>
<tr>
<td>Survival</td>
<td>46</td>
</tr>
<tr>
<td>Feeding and nutrition</td>
<td>45</td>
</tr>
<tr>
<td>The NICU^a stay</td>
<td>40</td>
</tr>
<tr>
<td>Quality of life</td>
<td>36</td>
</tr>
<tr>
<td>Outcomes</td>
<td>28</td>
</tr>
<tr>
<td>Emotional impact</td>
<td>26</td>
</tr>
</tbody>
</table>

^aNICU: neonatal intensive care unit.
Research conducted so far on AI chat engines has focused on maintaining the flow of conversation. The virtual patient remains free of jargon and her responses are largely focused on ensuring counseling. Modifications made to responses to make them more variation in realism and usability compared to VSPs may be beneficial in prenatal counseling education. There are less usable for a VSP that needs to maintain flow across many different variations of the same conversation. Only 2% (5/254) of the AI-generated responses were not usable in the VSP. Examples of minimally usable responses included “How much should I feed my baby each time?” which is not relevant to how feeding is done in the neonatal intensive care unit and “I am,” as this response is too nonspecific to be of use in a VSP. Of the 34% (87/254) of responses that required moderate modifications, the changes primarily involved adjusting terminology to ensure the parent was using colloquial, jargon-free language. As an example, “I’ve been having a lot of contractions and I’m only 24 weeks pregnant” was modified to “I’ve been having a lot of cramping and only 6 months pregnant.” Other modifications included adding some specificity to a response to ensure the VSP can use the sentence in the right context, such as modifying “That sounds reassuring, but what are the risks?” to “That sounds reassuring, but what are the risks of being born this early?” Of the 13% (34/254) of responses that required minimal adjustment, example changes included “I don’t trust the doctors” to “I don’t trust doctors,” and “Okay, thank you, but can you tell me more about what might happen to my baby in the future?” to “Can you tell me more about what might happen to my baby in the future?”

**Discussion**

**Principal Findings**

In this study, we examined the feasibility of using ChatGPT to enhance the realism of the VANESSA VSP. We found that the integration of ChatGPT generated many realistic, relevant, and useful responses. Based on these findings, ChatGPT-enabled VSPs may be beneficial in prenatal counseling education. There was more variation in realism and usability compared to relevance; therefore, an expert review was necessary to provide quality control before integrating the ChatGPT-generated conversations into an educational VSP program for prenatal counseling. Modifications made to responses to make them usable for the VANESSA VPS were largely focused on ensuring the virtual patient remains free of jargon and her responses maintain the flow of conversation.

Research conducted so far on AI chat engines has focused on using chat-based AI for the creation of discharge summaries, generating and interpreting electronic health records, assisting in medical education related to the medical licensing exam, and summarizing collections of journal articles to construct a brief abstract from the conclusions of the research [20-23]. The field is still relatively new, but rapidly increasing and expanding. This growth will only continue, as generating documentation and interacting with patients are key requirements of the health care setting. Health care simulation has many training applications, such as VSPs, that require expert authoring to educate clinicians and care providers on a certain skill or cognitive task. VSPs like VANESSA have been used in teaching the communication of medical ambiguity, evaluating medical students’ competence in performing critical clinical skills, and training nurses to recognize postpartum mood disorders [24-26]. Based on the results of our study, chat-based AI may be a valuable teaching tool in the future of health care simulation technology, leading to improved scenario creation, customization of patient interactions, and responses to care providers in a simulated setting. These improvements will result in authentic, unique interactive experiences, varying for each learner and training scenario.

We found that ChatGPT could generate many realistic parent responses, especially concerning issues related to survival at 24 weeks gestation and the neonatal intensive care unit stay (Figure S1 in Multimedia Appendix 1). Mistrust in the health care system is often encountered during stressful counseling conversations, and building the skill of responding to mistrust is crucial for physicians during their training [27]. Patients who express mistrust are less likely to engage with their health care team and care plan, and care is needed to proactively build trust during prenatal counseling [28-31]. Including opportunities for learners to respond to VSPs that express mistrust is one way to address this important counseling element, and ChatGPT provided a reliable mechanism to generate these phrases. Interestingly, the ChatGPT bot faced more challenges in generating realistic questions and responses about the emotional impact of preterm delivery and feeding. As these are frequently encountered topics of conversation in prenatal counseling, an expert review of these conversational elements remains a vital step before including them in an educational program.

ChatGPT produced responses that seemed relevant and appropriate to the context of prenatal counseling. Previous studies of prenatal counseling for extreme prematurity indicate that parents may ask questions about the likelihood of various outcomes, express a range of emotions, request engagement in shared decision-making, and express their parental roles and values [32,33]. Parents may express statements about their...

---

**Table 4.** Ratings of relevance, realism, and usability of sentences generated by ChatGPT (N=254).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Rating, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realism in parental responses and questions</td>
<td>1 (least) 8 (3) 38 (15) 20 (8) 183 (72)</td>
</tr>
<tr>
<td>Relevant to a prenatal counseling conversation</td>
<td>2 (1) 2 (1) 29 (11) 5 (2) 216 (85)</td>
</tr>
<tr>
<td>Usable for VSP(^a) educational program</td>
<td>5 (2) 1 (0) 87 (34) 34 (13) 127 (50)</td>
</tr>
</tbody>
</table>

\(^a\)VSP: virtual standardized patient.
uncertainty, anxieties, and hope for the future [34]. This wide range of topics, emotions, and questions makes it challenging to ensure that chatbot-generated conversations remain appropriate to the educational goals of the VSP. Despite the risk of getting off-topic, we found that only 1% (2/254) of ChatGPT-produced responses were irrelevant to a counseling conversation, given a carefully worded role-play prompt. Although most responses were relevant, some topics, such as spirituality and shared decision-making, did not come up in the role-play conversations. Previous studies have demonstrated that providers perceive the importance of parents’ spirituality in their decision-making and infrequently discuss these spiritual beliefs with parents in antenatal consultations [35,36]. Further work exploring how families might express their spiritual or explore shared decisions would be needed to ensure these topics are included in a VSP [37-39].

Chatbot programs use machine learning to generate their responses; due to the nature of machine learning, there is an inherent risk that chatbots can generate factually incorrect information [40]. Given this risk, caution is warranted when using chatbots in health care settings, where misinformation can have a significant risk [41,42]. Developers are working to address these inaccuracies as they design the next generation of large language model chat programs; they have demonstrated improvements in ChatGPT-4’s success across a variety of standardized tests [43]. This study leverages the strengths of a natural language chatbot in its ability to generate conversation while avoiding the risks of obtaining inaccurate medical information. Most scripts created by ChatGPT were usable for our perinatal counseling virtual patient. We found about a third of chatbot-generated phrases needed modification before being able to be integrated into a VSP; therefore, it may not be feasible to directly use ChatGPT for educational role-play without having the quality control step of review by expert clinicians. However, as technology continues to grow, this will evolve, and each subsequent model should be evaluated for usability.

Study Limitations

This exploratory study has several limitations. First, the pilot was done using ChatGPT 3.5, which is a single platform and is not representative of all chatbots. Later versions of ChatGPT have already been released and may have differences in realism, appropriateness, and usability. Newer AI chatbot programs are being trained on more parameters (175 billion for ChatGPT-3 vs an anticipated 100 trillion with ChatGPT-4), are supposed to have more ability to iterate on the same topic, and are being adjusted to improve the faculty accuracy of their responses [43]. Second, chatbot programs have limited information on which they build a conversation. For this study, we used a stable prompt around an impending 24-week gestation delivery to fit the standardized patient scenario, but conversations may be different with variations in the prompt. The AI was given a limited background to build a role-play, potentially limiting the diversity of ways in which patients could communicate their concerns. For this scenario, we requested a fifth-grade reading level for all patient roles to better mimic how patients may speak in stressful situations, but we did not explore higher or lower complexity of responses. Future work should explore how variations in the background, scenario, and reading level provided to the chatbot impact the output of the role-play. Another significant limitation was that response checking was performed by neonatologists, without input from families or trainees. Future work to refine the model will incorporate their views to ensure further applicability of the VSP and the validity of any assessments. Finally, although individual phrases exhibited good realism, the total duration of each patient-physician conversation (averaging 5 volleys) was generally shorter than that of a real prenatal counseling conversation.

Conclusions

Generative AI programs, such as ChatGPT, may provide a viable source of training materials to expand VSP programs with careful attention to the concerns and questions of patients and families. Given the potential for unrealistic or inappropriate statements and questions, an expert should review AI chat outputs before deploying them in an educational program.


Abbreviations

- **AI**: artificial intelligence
- **VANESSA**: Virtual Antenatal Encounter and Standardized Simulation Assessment
- **VSP**: virtual standardized patient
Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study

Annika Meyer¹; Janik Riese², BSc; Thomas Streichert¹, Prof Dr

¹Institute for Clinical Chemistry, University Hospital Cologne, Cologne, Germany
²Department of General Surgery, Visceral, Thoracic and Vascular Surgery, University Hospital Greifswald, Greifswald, Germany

Corresponding Author:
Annika Meyer
Institute for Clinical Chemistry
University Hospital Cologne
Kerpener Str 62
Cologne, 50937
Germany
Email: annika.meyer1@uk-koeln.de

Abstract

Background: The potential of artificial intelligence (AI)–based large language models, such as ChatGPT, has gained significant attention in the medical field. This enthusiasm is driven not only by recent breakthroughs and improved accessibility, but also by the prospect of democratizing medical knowledge and promoting equitable health care. However, the performance of ChatGPT is substantially influenced by the input language, and given the growing public trust in this AI tool compared to that in traditional sources of information, investigating its medical accuracy across different languages is of particular importance.

Objective: This study aimed to compare the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination.

Methods: To assess GPT-3.5’s and GPT-4’s medical proficiency, we used 937 original multiple-choice questions from 3 written German medical licensing examinations in October 2021, April 2022, and October 2022.

Results: GPT-4 achieved an average score of 85% and ranked in the 92.8th, 99.5th, and 92.6th percentiles among medical students who took the same examinations in October 2021, April 2022, and October 2022, respectively. This represents a substantial improvement of 27% compared to GPT-3.5, which only passed 1 out of the 3 examinations. While GPT-3.5 performed well in psychiatry questions, GPT-4 exhibited strengths in internal medicine and surgery but showed weakness in academic research.

Conclusions: The study results highlight ChatGPT’s remarkable improvement from moderate (GPT-3.5) to high competency (GPT-4) in answering medical licensing examination questions in German. While GPT-4’s predecessor (GPT-3.5) was imprecise and inconsistent, it demonstrates considerable potential to improve medical education and patient care, provided that medically trained users critically evaluate its results. As the replacement of search engines by AI tools seems possible in the future, further studies with nonprofessional questions are needed to assess the safety and accuracy of ChatGPT for the general population.

(JMIR Med Educ 2024;10:e50965) doi:10.2196/50965

KEYWORDS
ChatGPT; artificial intelligence; large language model; medical exams; medical examinations; medical education; LLM; public trust; trust; medical accuracy; licensing exam; licensing examination; improvement; patient care; general population; licensure examination

Introduction

Rapid advancements in large language models (LLMs) have sparked considerable excitement regarding their potential applications in the medical field [1,2]. One LLM-based application that has garnered worldwide attention is ChatGPT, developed by the research and deployment company OpenAI, due to its easy accessibility and potential to democratize knowledge [3]. The freely available version is based on the artificial intelligence (AI)–based tool GPT-3.5, which
encompasses billions of parameters and has been trained on approximately 570 GB of text from the internet [1,2].

ChatGPT’s GPT-3.5 iteration has already shown promise in several routine medical tasks and medical research [4-7], even raising ethical concerns in the literature [2,3,8]. The prompt and interactive nature of this AI’s responses might even revolutionize search engines, while also revealing shortcomings in medical education [9-11]. However, despite the introduction of the more advanced iteration GPT-4, concerns about the lack of transparency regarding this AI’s model parameters, training process, and underlying data structure remain unaddressed [8,12]. These concerns cast doubt on the medical proficiency of these LLMs, as both were not primarily trained on medical data and are the first to admit that as a language AI model, passing a medical examination is outside their skillset (Multimedia Appendix 1). Still, with assistance and adaptations, GPT-3.5 nearly passed the United States Medical Licensing Examination [13,14], and GPT-4 passed a Japanese medical examination [15]. Considering the variable performance of multilingual LLMs across different input languages [16,17], it is imperative to evaluate these models in various other linguistic contexts as well as on large data sets of original medical examination questions.

The primary objective of this study is to evaluate the medical proficiency of both ChatGPT iterations (GPT-3.5 and -4) in comparison to medical students by testing it on 937 original questions from the written German medical licensing examination (Zweites Staatsexamen), providing further data for a possible future integration. While the German medical licensing examination covers various medical subdisciplines in 320 multiple-choice questions [18], it has a high interexamination reliability of over 0.9 [19]. Despite using the same third-party client for question retrieval as earlier studies, the German approach of publicly releasing the examination questions enables the third-party client to guarantee the originality of the test items derived directly from the examination itself [20]. Additionally, to the best of our knowledge, we have tested both ChatGPT versions on the largest data set of medical licensing examination questions not included in their training data set. Furthermore, we did not exclude all image-based questions a priori. Instead, we evaluated the relevance of the images for each question and compared the results both with and without images.

**Methods**

**Data Collection**

To ensure that any observed performance was not influenced by changes in ChatGPT’s training data, we specifically chose the 3 most recent examinations (October 2021, April 2022, and October 2022) after the AI’s knowledge cutoff date [17]. Thus, we were able to obtain 937 multiple-choice questions, each with 5 possible answers from the third-party client Amboss, a web-based learning platform that provides the original questions from the Institut für Medizinische und Pharmazeutische Prüfungsfragen (IMPP). To maintain the original examination format, we presented all obtained questions and answer options in the same order as they appeared in the examination. No specific training code was used while submitting the questions. Due to AI’s inability to analyze visual content, answerability based on question text alone was defined as the primary inclusion criterion, resulting in the exclusion of 102 questions. The questions were submitted through ChatGPT’s interface of the GPT-3.5 (January 30, 2023) and GPT-4 (March 14, 2023) versions. ChatGPT’s answers were then compared to the official correct answers and evaluated. If ChatGPT selected none or more than 1 of the multiple-choice answers, the question was repeated in its original format up to 4 times or until a conclusive response could be obtained from ChatGPT (Figure 1).

We recorded additional data, such as answer length, content warnings, and recommendations for further diagnosis, and categorized the questioning methodology. To assess the readability of a question, we used the Simple Measure of Gobbledygook (SMOG) as it has shown acceptable interrater reliability for patient education materials in the literature [21]. Examination statistics provided by the “MEDI-LEARN” portal were also used, including the number of correct student answers and the specialization of each question. The “Blueprint” published by the IMPP outlines the distribution of subspecialties within the written state examinations [18].
Figure 1. Flowchart of the study design for the evaluation of ChatGPT’s (GPT-3.5 and GPT-3) accuracy in the written German medical licensing examination (2021-2022). The flowchart presents the criteria for question selection, including both the inclusion and exclusion criteria.

Statistical Analysis
To perform our data analysis, we used several packages [22-37] in addition to the R programming language [38].

While continuous variables were reported as arithmetic mean (SD) values, categorical variables were reported as frequencies and percentages. The Kolmogorov-Smirnov test, Shapiro-Wilk test, and QQ plots were used to confirm the normal distribution of continuous data statistically and graphically. To determine significant differences, we used unpaired t test or ANOVA for continuous variables and chi-square test or Wilcoxon rank-sum test for categorical variables. P values of <.05 were deemed significant. Univariate and multivariate regression analyses were additionally performed to provide information on probabilities and predictors.

Ethical Considerations
Ethics approval was not required as data were collected from publicly available sources on the internet or were generated using AI-based methods. No personally identifiable information was used in the data collection, and all data were handled in accordance with applicable data privacy laws and regulations.

Results
Overall, GPT-4 demonstrated superior performance with an average score of 796 out of 937 (85%), surpassing GPT-3.5’s score of 548 out of 937 (58%), which previously fell below the general passing threshold of 60% (Figure 2A) [37-39]. For the April 2022 examination, GPT-3.5 and GPT-4 achieved their highest scores (GPT-3.5: 195/319, 61%; GPT-4: 287/315, 91%), while the proportion of students who answered correctly remained constant across the 3 examinations (mean 76%, SD 18%; P=.86; Figure 2B and Multimedia Appendix 2).

Thus, GPT-4 passed all tested examinations, whereas GPT-3.5 could only pass 1 of the 3 examinations. Although the examinations varied in several aspects, we also observed a significant difference in the number of images (P=.02; Figure 2C and Multimedia Appendix 2). As GPT-3.5 and GPT-4 could, at the time of the study, not process these, we further investigated the potential image-related discrepancy between the examinations by excluding from subsequent analyses any questions that required image-dependent responses. The exclusion of these questions did not significantly alter examination difficulty, as evidenced by similar student scores (Figure 2D).

Moreover, no differences were observed in the parameters collected on student accuracy, questions, or answer characteristics in relation to the performance of GPT-4 and GPT-3.5 in the excluded cases (Multimedia Appendix 3). Upon excluding image-based questions, GPT-4 continued to outperform GPT-3.5, with scores approaching 91.44%. However, GPT-3.5 exceeded expectations by achieving passing scores on all 3 examinations (October 2021: 60.22%; April 2022: 63.36%; October 2022: 60.07%; Figure 2E and Multimedia Appendix 4). GPT-3.5’s accuracy (P=.66), the number of images (P=.07), and students’ accuracy (P=.77) remained constant throughout the examinations, whereas GPT-4’s accuracy (P=.02), the specialties (P<.001), and question type (P=.04) varied (Multimedia Appendix 4 and Figures 2A, 2B, and 2E). The details of the included questions and their respective categorizations are provided in Table 1.
Figure 2. Bar plots of ChatGPT’s (GPT-3.5 and GPT-4) and box plots of students’ accuracy in the written German medical licensing examination (2021-2022). Bar graphs and box plots of (A) the relative number of correct answers provided by ChatGPT (GPT-3.5 and GPT-4) answers, (B) correct answers provided by students, (C) and image-based questions for the different examinations. (D and E) The relative number of correct answers by ChatGPT (GPT-3.5 and GPT-4) and students, comparing all questions with the included text-based questions. The 60% pass mark is presented as a red line in (A) and (E) to provide context for the performance of ChatGPT (GPT-3.5 and GPT-4). In addition, (E) displays the percentile achieved by ChatGPT (GPT-3.5 and GPT-4) for each year’s examination, based on the percentile limits published by the Institut für Medizinische und Pharmazeutische Prüfungsfragen [37-39].
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Overall (N=834)</th>
<th>Accuracy of GPT-3.5</th>
<th>Accuracy of GPT-4</th>
<th>P value</th>
<th>Accuracy of GPT-3.5</th>
<th>Accuracy of GPT-4</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False (n=532)</td>
<td>True (n=302)</td>
<td>False (n=105)</td>
<td>True (n=729)</td>
<td>False (n=532)</td>
<td>True (n=302)</td>
<td>True (n=729)</td>
</tr>
<tr>
<td>Students' correct response rate (%)</td>
<td>77 (18)</td>
<td>71 (18)</td>
<td>80 (16)</td>
<td>&lt;.001&lt;sup&gt;a&lt;/sup&gt;</td>
<td>70 (18)</td>
<td>78 (17)</td>
<td>&lt;.001&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Accuracy of GPT-3.5, n (%)</td>
<td>511 (61)</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>N/A</td>
<td>N/A</td>
<td>38 (36)</td>
<td>473 (65)</td>
<td>&lt;.001&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Accuracy of GPT-4, n (%)</td>
<td>729 (87)</td>
<td>256 (79)</td>
<td>473 (93)</td>
<td>&lt;.001&lt;sup&gt;c&lt;/sup&gt;</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Readability score of the question, mean (SD)</td>
<td>14.96 (1.89)</td>
<td>14.93 (1.87)</td>
<td>14.98 (1.90)</td>
<td>.65&lt;sup&gt;a&lt;/sup&gt;</td>
<td>14.91 (2.26)</td>
<td>14.97 (1.84)</td>
<td>.21&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Question type, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connected (key feature)</td>
<td>532 (64)</td>
<td>204 (63)</td>
<td>328 (64)</td>
<td></td>
<td>79 (75)</td>
<td>453 (62)</td>
<td></td>
</tr>
<tr>
<td>Single question</td>
<td>302 (36)</td>
<td>119 (37)</td>
<td>183 (36)</td>
<td>.02&lt;sup&gt;c&lt;/sup&gt;</td>
<td>26 (25)</td>
<td>276 (38)</td>
<td>.03&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Images referenced in questions</td>
<td>84 (10)</td>
<td>23 (7.1)</td>
<td>61 (12)</td>
<td></td>
<td>17 (16)</td>
<td>67 (9.2)</td>
<td></td>
</tr>
<tr>
<td>Specialty, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gynecology</td>
<td>43 (5.2)</td>
<td>12 (3.7)</td>
<td>31 (6.1)</td>
<td></td>
<td>7 (6.7)</td>
<td>36 (4.9)</td>
<td></td>
</tr>
<tr>
<td>Infectiology</td>
<td>74 (8.9)</td>
<td>24 (7.4)</td>
<td>50 (9.8)</td>
<td></td>
<td>6 (5.7)</td>
<td>68 (9.3)</td>
<td></td>
</tr>
<tr>
<td>Internal medicine</td>
<td>176 (21)</td>
<td>71 (22)</td>
<td>105 (21)</td>
<td></td>
<td>15 (14)</td>
<td>161 (22)</td>
<td></td>
</tr>
<tr>
<td>Neurology</td>
<td>112 (13)</td>
<td>51 (16)</td>
<td>61 (12)</td>
<td></td>
<td>12 (11)</td>
<td>100 (14)</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>269 (32)</td>
<td>106 (33)</td>
<td>163 (32)</td>
<td></td>
<td>46 (44)</td>
<td>223 (31)</td>
<td></td>
</tr>
<tr>
<td>Pediatrics</td>
<td>62 (7.4)</td>
<td>26 (8.0)</td>
<td>36 (7.0)</td>
<td>.02&lt;sup&gt;c&lt;/sup&gt;</td>
<td>11 (10)</td>
<td>51 (7.0)</td>
<td>.03&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>54 (6.5)</td>
<td>11 (3.4)</td>
<td>43 (8.4)</td>
<td></td>
<td>5 (4.8)</td>
<td>49 (6.7)</td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td>44 (5.3)</td>
<td>22 (6.8)</td>
<td>22 (4.3)</td>
<td>.64&lt;sup&gt;c&lt;/sup&gt;</td>
<td>3 (2.9)</td>
<td>41 (5.6)</td>
<td>.34&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Expertise, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background knowledge</td>
<td>103 (12)</td>
<td>32 (9.9)</td>
<td>71 (14)</td>
<td></td>
<td>13 (12)</td>
<td>90 (12)</td>
<td></td>
</tr>
<tr>
<td>Complications</td>
<td>49 (5.9)</td>
<td>19 (5.9)</td>
<td>30 (5.9)</td>
<td></td>
<td>4 (3.8)</td>
<td>45 (6.2)</td>
<td></td>
</tr>
<tr>
<td>Diagnostic competence</td>
<td>466 (56)</td>
<td>184 (57)</td>
<td>282 (55)</td>
<td></td>
<td>54 (51)</td>
<td>412 (57)</td>
<td></td>
</tr>
<tr>
<td>Prevention competence</td>
<td>36 (4.3)</td>
<td>13 (4.0)</td>
<td>23 (4.5)</td>
<td></td>
<td>6 (5.7)</td>
<td>30 (4.1)</td>
<td></td>
</tr>
<tr>
<td>Scientific practice</td>
<td>34 (4.1)</td>
<td>14 (4.3)</td>
<td>20 (3.9)</td>
<td></td>
<td>8 (7.6)</td>
<td>26 (3.6)</td>
<td></td>
</tr>
<tr>
<td>Therapeutic competence</td>
<td>146 (18)</td>
<td>61 (19)</td>
<td>85 (17)</td>
<td></td>
<td>20 (19)</td>
<td>126 (17)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Wilcoxon rank-sum test.
<sup>b</sup>N/A: not applicable.
<sup>c</sup>Pearson chi-square test.

After controlling for all other variables, correct student responses (GPT-3.5: OR 0.01, 95% CI 0.00-0.01, P<.001; GPT-4: OR 0.00, 95% CI 0.00-0.00, P=.003) and questions with images (GPT-3.5: OR 0.19, 95% CI 0.08-0.30, P<.001; GPT-4: OR –0.09, 95% CI –0.16 to –0.01, P=.02) emerged as significant predictors of GPT-3.5’s and GPT-4’s accuracy, regardless of the version. Furthermore, our analysis revealed that only questions pertaining to psychiatry were significant predictors of correct GPT-3.5 responses (OR 0.19, 95% CI 0.02-0.36, P=.03). In contrast, questions related to internal medicine (OR 0.10, 95% CI 0.00-0.19, P=.04) and surgery (OR 0.12, 95% CI 0.00-0.25, P=.049) were the only medical subspecialties significantly predicting accurate responses of GPT-4. Conversely, questions concerning scientific practice (OR –0.14, 95% CI –0.29 to 0.00, P=.05) were less likely to be answered correctly by GPT-4 (Table 2 and Figure 3). The question SMOG readability score, however, did not significantly impact ChatGPT’s accuracy.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>GPT-3.5</th>
<th>GPT-4</th>
<th>GPT-3.5</th>
<th>GPT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate</td>
<td>Univariate</td>
<td>Multivariate</td>
<td>Multivariate</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td>95% CI</td>
<td>P value</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>Students' correct response rate</td>
<td>1.03</td>
<td>1.02 to 1.04</td>
<td>&lt;.001</td>
<td>.01</td>
</tr>
<tr>
<td>Accuracy of GPT-4</td>
<td>3.25</td>
<td>2.13 to 5.02</td>
<td>&lt;.001</td>
<td>.26</td>
</tr>
<tr>
<td>Accuracy of GPT-3.5</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>October 2021 examination</td>
<td>0.94</td>
<td>0.70 to 1.27</td>
<td>.68</td>
<td>.00</td>
</tr>
<tr>
<td>April 2022 examination</td>
<td>1.15</td>
<td>0.86 to 1.54</td>
<td>.35</td>
<td>.03</td>
</tr>
<tr>
<td>October 2022 examination</td>
<td>0.92</td>
<td>0.69 to 1.24</td>
<td>.59</td>
<td>N/A</td>
</tr>
<tr>
<td>Question type</td>
<td>0.96</td>
<td>0.72 to 1.28</td>
<td>.78</td>
<td>–.03</td>
</tr>
<tr>
<td>Images referenced in questions</td>
<td>1.77</td>
<td>1.09 to 2.98</td>
<td>.03</td>
<td>.19</td>
</tr>
<tr>
<td>Other specialty</td>
<td>0.96</td>
<td>0.71 to 1.30</td>
<td>.80</td>
<td>.00</td>
</tr>
<tr>
<td>Gynecology and obstetrics</td>
<td>1.62</td>
<td>0.84 to 3.33</td>
<td>.17</td>
<td>.12</td>
</tr>
<tr>
<td>Surgery</td>
<td>0.62</td>
<td>0.33 to 1.14</td>
<td>.12</td>
<td>–.12</td>
</tr>
<tr>
<td>Internal medicine</td>
<td>0.92</td>
<td>0.66 to 1.30</td>
<td>.63</td>
<td>–.02</td>
</tr>
<tr>
<td>Infectious diseases</td>
<td>1.35</td>
<td>0.82 to 2.28</td>
<td>.24</td>
<td>.06</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>2.61</td>
<td>1.37 to 5.40</td>
<td>.005</td>
<td>.19</td>
</tr>
<tr>
<td>Neurology</td>
<td>0.72</td>
<td>0.49 to 1.08</td>
<td>.12</td>
<td>–.04</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>0.87</td>
<td>0.52 to 1.48</td>
<td>.60</td>
<td>N/A</td>
</tr>
<tr>
<td>Diagnostic competence</td>
<td>0.93</td>
<td>0.70 to 1.23</td>
<td>.60</td>
<td>–.03</td>
</tr>
<tr>
<td>Therapeutic competence</td>
<td>0.86</td>
<td>0.60 to 1.24</td>
<td>.41</td>
<td>–.04</td>
</tr>
<tr>
<td>Background knowledge</td>
<td>1.47</td>
<td>0.95 to 2.32</td>
<td>.09</td>
<td>.08</td>
</tr>
<tr>
<td>Prevention competence</td>
<td>1.13</td>
<td>0.57 to 2.32</td>
<td>.74</td>
<td>.00</td>
</tr>
<tr>
<td>Scientific practice</td>
<td>0.90</td>
<td>0.45 to 1.85</td>
<td>.77</td>
<td>.01</td>
</tr>
<tr>
<td>Complications</td>
<td>1.00</td>
<td>0.56 to 1.84</td>
<td>&gt;.99</td>
<td>N/A</td>
</tr>
<tr>
<td>Readability score of the question</td>
<td>1.01</td>
<td>0.94 to 1.09</td>
<td>.70</td>
<td>.01</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
**Discussion**

**Principal Findings**

With the introduction of ChatGPT’s GPT-3.5 and GPT-4 iterations, the potential application for AI in research, patient care, and medical education is gaining recognition [2,8,40]. By improving the users’ experience and facilitating more efficient information retrieval, ChatGPT might even revolutionize the future of search engines and shift the focus of medical education from memorization to practical application [8,10,11].

Under this premise, the nearly passing scores of the freely available GPT-3.5 iteration, along with the exceptional scores of GPT-4, are highly relevant. Even with the varying scores of 51%-67% of GPT-3.5 across various input languages [13-15,41,42], both models consistently outperform most prominent general and domain-specific LLMs, such as
InstructGPT (53%), GPT-3 (25%), and BioMedLM (50%) [14,43,44]. Despite these improvements, GPT-3.5’s or GPT-4’s performance still fell short in comparison to that of medical students in a Japanese medical examination according to the study by Takagi et al [15]. In comparison to the German medical students, however, GPT-3.5 scored in the 8.6th percentile, while GPT-4 ranked in the 92.8th, 99.5th, and 92.6th percentiles in the October 2021, April 2022, and October 2022 examinations [39,45,46]. The observed variations in the AI’s accuracy across input languages may partially reflect the language composition of their data sets, as LLMs tend to favor languages that are more represented in their training data [16,17]. Since ChatGPT appears to perform optimally with English inputs, language emerges as a limiting factor for its accuracy, suggesting that globally consistent application is dependent upon users’ proficiency in English.

Moreover, the nearly 30% performance increase from GPT-3.5 to GPT-4, as indicated in this study and supported by a Japanese study, which suggests a similar language distribution within the GPT-3.5 and GPT-4 data sets [15]. GPT-4, unlike GPT-3.5, also did not answer questions containing images on repetition, showing an improvement in the previously incorrect content produced by GPT-4’s predecessor [17].

Thus, health care professionals could potentially benefit, especially from GPT-4’s conclusive and often nonobvious insights to multiple-choice questions, as these users have the ability to verify crucial details [13,14,41]. For instance, there is potential for using GPT-3.5 and GPT-4 in a medical education tutoring environment, as evidenced by its successful application in anatomy [47]. However, when using either GPT-3.5 or GPT-4 for medical applications, its differing accuracy across specialties must also be taken into account [48]. GPT-3.5 initially displayed a higher degree of accuracy within the field of psychiatry, while GPT-4 demonstrated its strength in internal medicine and surgery. Considering the rising prevalence of psychiatric disorders and concomitant challenges in providing care, it seems likely that nonprofessionals would also turn to the chatbot for mental health issues at the time of GPT-3.5’s release [8,49,50]. Hence, it is conceivable that GPT-3.5’s training data set includes not only a substantial and reliable portion of psychiatric data, but also its developers might have first fine-tuned ChatGPT specifically in this domain in anticipation of its high demand [51-53]. Thus, the developers might have also fine-tuned GPT-4 specifically in internal medicine and surgery, possibly reacting to a high demand in this area from users of its predecessor. GPT-4’s impressive performance is not limited to the medical field, as it demonstrated comparable percentile scores in the Uniform Bar Exam, showcasing its potential as a versatile tool across diverse academic disciplines [17]. However, assessing the possible reasons for the performance differences between GPT-3.5 and GPT-4 is complicated by the confidential architecture of GPT-4 [54], posing challenges for research on future applications.

In turn, GPT-4’s excellent achievements shed light on the limitations of current testing paradigms in medical education that often favor rote memorization over a critical and context-aware approach. They also highlight the inadequacy of multiple-choice questions as a means of assessing medical knowledge, as they tend to encourage binary thinking as “true” and “false,” which often fails to capture the complex reality of the medical practice [11]. Although GPT-3.5 and GPT-4 allow the simple and fast retrieval of medical information from any internet-capable device that fits in one’s pocket [9,10], neither GPT-3.5 nor GPT-4 verifies the information they provide. Thus, ChatGPT’s output needs to be approached with a critical mindset, recognizing that misinformation may be more difficult to detect than in the output of other search engines that offer multiple sources in response to a query and take login credentials into account [8,55]. To navigate these changing informational landscapes, a basic understanding in data science seems necessary alongside traditional medical expertise [56]. It may even be beneficial for future iterations of AI tools to include references to the sources underlying each search in order to increase transparency and allow users to assess the reliability of the information they receive.

In a previous study by Nov et al [57], considering that 59% of participants trusted chatbots more than traditional search engines, it must be noted that GPT-3.5 and GPT-4 have only been tested on medical examination questions and not questions by nonprofessionals, limiting general recommendations for unsupervised patient education or the general population. It seems evident that GPT-4 has been benchmarked against medical licensing examinations, explaining not only GPT-4’s excellent scores but also exceeding achievements in internal medicine and surgery, which, for instance, have been overrepresented in the medical examinations assessed in this study [12,17].

Since GPT-3.5 failed the German medical licensing examination by a narrow margin, its use for answering medical questions is generally not advisable. Moreover, the remarkable performance of GPT-4 in the German Medical State Examination may not be universally applicable outside a medical examination setting, especially considering that GPT-4 was presumably benchmarked on academic and professional examinations [17].

As literature on ChatGPT is scarce, and it can be difficult to detect incorrect output from this AI tool, the content it generates must be carefully assessed. Nevertheless, medical professionals may still be able to benefit from GPT-3.5’s and GPT-4’s explanations and, in some cases, gain new nonobvious insights. With the release of GPT-4’s ability to handle pictures on the horizon, the potential for further applications of GPT-3.5 and GPT-4 to improve the medical workflow or medical education seems eminent, emphasizing the need for continued research into AI.

Limitations

This study’s findings on GPT-3.5’s and GPT-4’s medical proficiencies are limited to multiple-choice questions from the German medical licensing examination, which may not be representative of other types of examinations or contexts. However, it is worth noting that GPT-3.5 and GPT-4 have demonstrated similar performances in examinations in other countries and languages, which suggests some degree of generalizability.
In addition, the sample size of 937 questions and the exclusion of image-based questions may not capture the full range of difficulty levels or content areas. Although the collected parameters did not differ in terms of GPT-3.5’s and GPT-4’s accuracy in the excluded cases, the decision to exclude image-based questions may have introduced a sampling bias. By testing for differences, efforts were made to minimize this bias and maintain the integrity of the results.

As GPT-3.5’s and GPT-4’s performances were compared to those of German medical students using the MEDI-LEARN service, a selection bias might have been introduced. However, the high correlation between the MEDI-LEARN statistics and the IMPP statistics indicates at best a weak expression of this selection bias [58].

It should also be noted that a replication of this study might not yield the exact same results, as the literature suggests that GPT-3.5 is inconsistent in answering 15% of medical questions [59]. However, the trends observed in this study appear to be consistent with those reported in other published and preprint studies on GPT-3.5’s and GPT-4’s performance.

Conclusions

In conclusion, the results of this study indicate that only GPT-4 consistently passed all 3 medical examinations, ranking in the 92.8th to 99.5th percentile in comparison to medical students. These findings highlight the strengths and limitations of ChatGPT in the context of medical examinations and raise questions about the future of medical education.

Although GPT-3.5’s and GPT-4’s accuracy in medical examinations seems consistent across different countries and languages, its inconsistencies, potential biases, and number of incorrect answers restrain a recommendation for its use by the general population for medical purposes. However, its elaborate explanations and potential to yield nonobvious insights may benefit medical professionals in training.

While this study hints to a moderate accuracy of GPT-3.5 and a stellar performance of GPT-4 in answering medical examination questions, further research is necessary to gain deeper insights, explore future applications, and ensure safe use of ChatGPT for end users.

Acknowledgments

The authors thank Dorothee Meyer, Linea Luise Fuchs, Ari Soleman, GPT-3.5, and GPT-4 for proofreading this manuscript. In this study, we used ChatGPT for several purposes: to translate our manuscript into English, to refine its linguistic presentation, to evaluate and improve our methodological approach, and to scrutinize the R code underlying our statistical analysis, with a particular focus on identifying and resolving any error warnings generated. Subsequently, all outputs provided by ChatGPT were rigorously reviewed and critically appraised by the authors to ensure accuracy and reliability.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Responses of (A) GPT-3.5 and (B) GPT-4 to the queries on its ability to pass a medical exam, 2023.
[DOCX File, 592 KB - mededu_v10i1e50965_app1.docx]

Multimedia Appendix 2

Summary statistics for all questions regarding exam time and ChatGPT’s (GPT-3.5 and GPT-4) accuracy in the German medical licensing exam, 2021-2022.
[DOCX File, 21 KB - mededu_v10i1e50965_app2.docx]

Multimedia Appendix 3

Summary statistics for excluded questions regarding ChatGPT’s (GPT-3.5 and GPT-4) accuracy in the German medical licensing exam, 2021-2022.
[DOCX File, 20 KB - mededu_v10i1e50965_app3.docx]

Multimedia Appendix 4

Summary statistics for included questions regarding exam time in the German medical licensing exam, 2021-2022.
[DOCX File, 17 KB - mededu_v10i1e50965_app4.docx]

References


17. Examen (M2/M3) No.1 in der Examensvorbereitung. AMBOSS. URL: https://www.amboss.com/de/examen-m2-m3 [accessed 2023-07-15]


58. FAQ: Häufig gefrage Fragen. MEDI-LEARN. URL: https://www.mlmr.de/unis/faq/#faq1 [accessed 2024-01-28]


**Abbreviations**

AI: artificial intelligence  
IMPP: Institut für Medizinische und Pharmazeutische Prüfungsfragen  
LLM: large language model  
SMOG: Simple Measure of Gobbledygook

©Annika Meyer, Janik Riese, Thomas Streichert. Originally published in JMIR Medical Education (https://mededu.jmir.org), 08.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study

Peng Yu†, MD; Changchang Fang†, MD; Xiaolin Liu†, MD; Wanying Fu†, MD; Jitao Ling†, MD; Zhiwei Yan§, MD; Yuan Jiang∥, MD; Zhengyu Cao∥, MD; Maoxiong Wu∥, MD; Zhiteng Chen∥, MD; Wengen Zhu∥, MD; Yuling Zhang∥, MD; Ayiguli Abudukeremu∥, MD; Yue Wang∥, MD; Xiao Liu∥, MD; Jingfeng Wang∥, MD

1Department of Endocrine, The Second Affiliated Hospital of Nanchang University, Jiangxi, China
2Department of Cardiology, The Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, China
3College of Kinesiology, Shenyang Sport University, Shenyang, China
4Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China
5Department of Cardiology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
*these authors contributed equally

Corresponding Author:
Xiao Liu, MD
Department of Cardiology
107 Yanjiang West Road
Guangzhou
China
Phone: 86 15083827378
Email: liux587@mail.sysu.edu.cn

Abstract

Background: ChatGPT, an artificial intelligence (AI) based on large-scale language models, has sparked interest in the field of health care. Nonetheless, the capabilities of AI in text comprehension and generation are constrained by the quality and volume of available training data for a specific language, and the performance of AI across different languages requires further investigation. While AI harbors substantial potential in medicine, it is imperative to tackle challenges such as the formulation of clinical care standards; facilitating cultural transitions in medical education and practice; and managing ethical issues including data privacy, consent, and bias.

Objective: The study aimed to evaluate ChatGPT’s performance in processing Chinese Postgraduate Examination for Clinical Medicine questions, assess its clinical reasoning ability, investigate potential limitations with the Chinese language, and explore its potential as a valuable tool for medical professionals in the Chinese context.

Methods: A data set of Chinese Postgraduate Examination for Clinical Medicine questions was used to assess the effectiveness of ChatGPT’s (version 3.5) medical knowledge in the Chinese language, which has a data set of 165 medical questions that were divided into three categories: (1) common questions (n=90) assessing basic medical knowledge, (2) case analysis questions (n=45) focusing on clinical decision-making through patient case evaluations, and (3) multichoice questions (n=30) requiring the selection of multiple correct answers. First of all, we assessed whether ChatGPT could meet the stringent cutoff score defined by the government agency, which requires a performance within the top 20% of candidates. Additionally, in our evaluation of ChatGPT’s performance on both original and encoded medical questions, 3 primary indicators were used: accuracy, concordance (which validates the answer), and the frequency of insights.

Results: Our evaluation revealed that ChatGPT scored 153.5 out of 300 for original questions in Chinese, which signifies the minimum score set to ensure that at least 20% more candidates pass than the enrollment quota. However, ChatGPT had low accuracy in answering open-ended medical questions, with only 31.5% total accuracy. The accuracy for common questions, multichoice questions, and case analysis questions was 42%, 37%, and 17%, respectively. ChatGPT achieved a 90% concordance across all questions. Among correct responses, the concordance was 100%, significantly exceeding that of incorrect responses (n=57, 50%; P<.001). ChatGPT provided innovative insights for 80% (n=132) of all questions, with an average of 2.95 insights per accurate response.
Conclusions: Although ChatGPT surpassed the passing threshold for the Chinese Postgraduate Examination for Clinical Medicine, its performance in answering open-ended medical questions was suboptimal. Nonetheless, ChatGPT exhibited high internal concordance and the ability to generate multiple insights in the Chinese language. Future research should investigate the language-based discrepancies in ChatGPT’s performance within the health care context.

(Keywords: ChatGPT; Chinese Postgraduate Examination for Clinical Medicine; medical student; performance; artificial intelligence; medical care; qualitative feedback; medical education; clinical decision-making)

Introduction
Artificial intelligence (AI) was initially conceptualized in 1956 [1], but it has only gained significant momentum in recent years. AI aims to replicate human intelligence and thinking processes through the use of brain-like computer systems to solve complex problems. What is most inspiring is that AI systems can be trained on specific data sets to improve prediction accuracy and tackle intricate problems [2-4], which means that one of the possible applications of AI is the ability to help doctors to rapidly search through vast amounts of medical data, enhancing their creativity and enabling them to make error-free decisions [5,6].

ChatGPT (OpenAI) is an AI model that has spurred great attention due to the revolutionary innovations in its ability to perform a diverse array of natural language tasks. By using a class of large-scale language models, ChatGPT (version 3.5) can predict the likelihood of a sequence of words based on the context of the preceding words. With sufficient training on vast amounts of text data, ChatGPT can generate novel word sequences that closely resemble natural human language and have never been observed before by other AI [7].

A study was conducted on the effectiveness of the version of generative pretrained transformer’s large-scale language model (ChatGPT, version 3.5) in passing the United States Medical Licensing Examination (USMLE). The results showed that the AI model achieved an accuracy rate of over 50% in all the tests, and in some analyses, it even surpassed 60% accuracy. It is imperative to highlight and emphasize that the study was conducted mostly using English input, and the AI model was also trained in English.

However, despite the success of AI models like ChatGPT in the English language, their performance in understanding and generating medical text in the Chinese language remains largely unexplored because ChatGPT’s ability to understand and generate text in any given language is limited by the quality and quantity of training data available in that language. Chinese is the second-most widely spoken language in the world, with more than 1.3 billion speakers globally, while the quality and quantity of Chinese language data may not be compared with English due to some reasons, such as complexity of the written characters. Thus, the performance of ChatGPT in Chinese medical information warrants further investigation.

In this study, ChatGPT’s clinical reasoning ability was evaluated by administering questions from the Chinese Postgraduate Examination for Clinical Medicine. This standardized and regulated test assesses candidates’ comprehensive abilities. The questions are textually and conceptually dense, and the difficulty and complexity of the questions are highly standardized and regulated. Additionally, this examination has demonstrated remarkable stability in raw scores and psychometric properties over the past years. Moreover, the examination comprises 43% (n=71) basic science and medical humanities, with 14% (n=23) physiology, 10% (n=17) biochemistry, 13% (n=28) pathology, and 6% (n=10) medical humanities. Clinical medicine makes up the remaining 57% (n=94), with internal medicine and surgery accounting for 37% (n=61) and 20% (n=33), respectively. Due to the examination’s linguistic and conceptual complexity, we hypothesize that it will serve as an excellent challenge for ChatGPT. By evaluating ChatGPT’s performance on this examination, we aimed to gain insights into the AI model’s potential for understanding and generating medical text in Chinese and assess its applicability in Chinese medical education and clinical practice.

Methods
Ethical Considerations
This study does not involve direct interaction with human participants or the collection of personal identifiable information. As a result, it falls under the category of nonhuman subject research. Therefore, no human subject ethics review approvals were required for this study. Since this study does not involve human participants or the collection of personal identifiable information, obtaining informed consent from individuals is not applicable. As this study does not involve the collection or use of personal identifiable information, privacy and confidentiality concerns are not applicable. Since this study does not involve human participants, there is no compensation provided to individuals.

Artificial Intelligence
ChatGPT uses self-attention mechanisms and extensive training data to generate contextually relevant responses in a conversational setting. It excels in managing long-range dependencies and creating coherent replies. However, it is important to clarify that ChatGPT (version 3.5), a server-based language model, does not possess internet browsing or search functionalities. Consequently, its responses are constructed solely on abstract relationships between words or “tokens” within its neural network [7]. Furthermore, it should be noted that OpenAI released the latest version, ChatGPT (version 4), in March 2023, but the data in this study were from February 2023, when ChatGPT (version 3.5) was the most recent version.
Input Source
The Chinese Postgraduate Examination for Clinical Medicine questions from 2022 were not officially released. However, a comprehensive set of 165 questions totaling 500 points was found on the web (Table S1 in Multimedia Appendix 1) and treated as original questions. Point values differed among question types: each case analysis question (CAQ) and multichoice question (MCQ) was worth 2 points, while common questions (CQs) were either worth 1.5 or 2 points each. All inputs fed into the ChatGPT (version 3.5) model were valid samples, not part of the training data set. This was due to the database not being updated since September 2021, predating the release of these questions. For future research convenience, the 165 questions were categorized into three types:

1. **CQs (n=90)**: These questions are to evaluate the knowledge of basic science in physiology, biochemistry, pathology, and medical humanities. Each question has 4 choices, and the respondent should select only the correct answers. For example: “Which can inhibit insulin secretion? (A) Increased free fatty acids in blood, (B) Increased gastric inhibitory peptide secretion, (C) Sympathetic nerve excitation, (D) Growth hormone secretion increases” was reformatted as “Can an increase in free fatty acids in the blood, an increase in gastric inhibitory peptide secretion, an increase in sympathetic nerve excitation, or an increase in growth hormone secretion inhibit insulin secretion?” This encoding strategy was applied across all 3 subgroups.

For instance, an original question like, “Which can inhibit insulin secretion? (A) Increased free fatty acids in blood, (B) Increased gastric inhibitory peptide secretion, (C) Sympathetic nerve excitation, (D) Growth hormone secretion increases” was reformatted as “Can an increase in free fatty acids in the blood, an increase in gastric inhibitory peptide secretion, an increase in sympathetic nerve excitation, or an increase in growth hormone secretion inhibit insulin secretion?” This encoding strategy was applied across all 3 subgroups.

Additionally, to mitigate potential memory retention bias, we commenced a new chat session for each query. This process of reformattling questions, presenting them to ChatGPT, and initiating new sessions for each question constituted our methodology for evaluating ChatGPT’s performance using the data set. The clarity of this process should address the concerns raised in the comment about the lack of understanding of the way we used the data set for evaluation.

2. **CAQs (n=45)**. It is a method used in clinical medicine to examine and evaluate patient cases. It involves an in-depth review of a patient’s medical history, presenting symptoms, laboratory and imaging results, and diagnostic findings to arrive at a diagnosis and treatment plan. There are 4 choices, and the respondent should select only the correct answers. The difference between CAQs and CQs is that CAQs focus on clinical decision-making. For example: “A 38-year-old male suffering chest pain and fever for 3 days, having a 5 years of diabetes history. Physical examination: t=37.6 °C, right lower lung turbid knock, breathing sound is reduced. A chest X radiograph suggests a right pleural effusion. Pleural aspiration liquefaction test showed WBC 650x106/L with fine lymph Cell 90% in pleural fluid, with glucose of 3.2 mmol/L, the diagnosis for this patient is? (A) Tuberculous pleurisy, (B) Malignant pleural effusion, (C) Empyema, (D) Pneumonia-like pleural effusion.”

3. **MCQs (n=30)**: There are 4 choices, and the respondent should select at least 2 correct answers. There is no point for choosing more or less. For example: “The structures of auditory bone conduction include? (A) Skull, (B) Round window film, (C) Ossicular chain, (D) Cochlear bone wall.”

Scoring
Initially, the question format had to be adjusted to properly evaluate the performance of ChatGPT in the Chinese Postgraduate Examination for Clinical Medicine questions. Specifically, we included a “multichoice” or “single-choice” notation, as we found ChatGPT’s responses varied without these cues. MCQs were adjusted to state “Please choose one or more correct options,” while CQs and CAQs were altered to indicate “There is only one correct answer.” This adjustment was necessary for evaluating ChatGPT’s performance in the Chinese language.

We then compiled a data set of these examination questions along with their correct answers. To ensure validity, the answers were cross-verified with web-based resources and consultations with senior doctors. ChatGPT’s performance was then evaluated by comparing its responses to the standard answers in the data set. A high examination score would suggest that ChatGPT handled this task effectively.

In our comprehensive analysis, we also delved into examining the correlation between different question types and accuracy using the Pearson correlation coefficient as a statistical measure to investigate this relationship.

Encoding
The structured examination questions were transformed into open-ended inquiries for better simulation of real-world clinical scenarios. Multiple-choice questions for the CAQ were removed, and ChatGPT was required to diagnose the patient’s disease and prove its reason.

Regarding the MCQs, we eliminated all the choices and did not prompt ChatGPT about the existence of multiple correct answers. The CQs were treated similarly to the MCQs. However, we encountered a distinct subset within these 3 categories that could not be processed like the other questions. This subset comprised questions that required 1 answer choice to be selected from the provided options. Therefore, these questions were converted into a special format (n=26).

For instance, an original question like, “Which can inhibit insulin secretion? (A) Increased free fatty acids in blood, (B) Increased gastric inhibitory peptide secretion, (C) Sympathetic nerve excitation, (D) Growth hormone secretion increases” was reformatted as “Can an increase in free fatty acids in the blood, an increase in gastric inhibitory peptide secretion, an increase in sympathetic nerve excitation, or an increase in growth hormone secretion inhibit insulin secretion?” This encoding strategy was applied across all 3 subgroups.

Additionally, to mitigate potential memory retention bias, we commenced a new chat session for each query. This process of reformattting questions, presenting them to ChatGPT, and initiating new sessions for each question constituted our methodology for evaluating ChatGPT’s performance using the data set. The clarity of this process should address the concerns raised in the comment about the lack of understanding of the way we used the data set for evaluation.

Adjudication
To assess ChatGPT’s performance thoroughly, 2 physicians independently scored AI outputs for accuracy, concordance, and insight using predefined criteria (Table S2 in Multimedia Appendix 1). These physicians were not aware of each other’s evaluations. To familiarize the physicians with the scoring system, a subset of 20 questions was used for training, during which the physicians were unblinded to each other’s assessments.

ChatGPT’s responses were classified into 3 categories under the accuracy parameter: accurate, inaccurate, and indeterminate. “Accurate” responses were those where ChatGPT provided the right answer, while “inaccurate” encompassed instances of no answer, an incorrect response, or multiple answers containing incorrect options. “Indeterminate” responses were those where...
the AI output did not present a definitive answer, suggesting insufficient information to make a selection.

Concordance was determined by whether ChatGPT’s explanation affirmed its provided answer, with discordance occurring if the explanation contradicted the answer. We defined valuable insights as unique text segments within the AI’s explanations meeting specific criteria: they were nondefinitional, nonobvious, valid, and unique. These insights required additional knowledge or deductions beyond the input question, provided accurate clinical or numerical information, and potentially eliminated multiple answer choices with a single insight.

To mitigate potential within-item anchoring bias, the adjudicators first evaluated the accuracy for all items, followed by concordance. In case of discrepancies in domain assessments, a third physician adjudicator was consulted. This third-party intervention was required for 11 items (n=11, 7% of the data set). We used the Cohen \( \kappa \) statistic to evaluate the interrater agreement between the physicians for the questions (Table S3 in Multimedia Appendix 1). A schematic overview of the study protocol is presented in Figure 1 to provide a clearer understanding of our methodology.

Figure 1. Schematic of workflow for sourcing, encoding, and adjudicating results. The 165 questions were categorized into 3 types: CQ, CAQ, and MCQ, and each question was assessed for its score. The accuracy of the CQ and MCQ questions was evaluated, while the MCQ questions were also assessed for the accuracy, concordance, and frequency of insights. The adjudication process was carried out by 2 physicians, and in case of any discrepancies in the domains, a third physician was consulted for adjudication. Additionally, any inappropriate output was identified and required re-encoding. CAQ: case analysis question; CQ: common question; MCQ: multichoice question.

Results
ChatGPT Performs Poor Toward the Original Questions

After inputting the original questions into ChatGPT and collecting their answers, ChatGPT received a score of 153.5 out of 300, which means that it only obtained 51.2% of the total points on the test. This score is much lower than expected but slightly higher than the passing threshold (129/300) defined by official agencies.

Among 3 subgroups of questions, the evaluation revealed that of a total of 90 CQs, ChatGPT only provided 50 (56%, 95% CI 45%–66%) correct answers. Similarly, of 45 CAQs, ChatGPT provided 25 (56%, 95% CI 41%–70%) correct answers. Furthermore, of 30 MCQs, ChatGPT provided 10 (33%, 95% CI 16%–50%) completely accurate answers (Figure 2). These results suggest that ChatGPT’s ability to resolve medical problems in Chinese needs to be improved.

Additionally, we have noticed a Pearson correlation coefficient value of approximately 0.228. This finding suggests a relatively weak correlation between the different question types and the accuracy of the responses.
ChatGPT Performs Worse on Encoded Questions Compared to the Original Questions

We encoded questions from the Chinese Postgraduate Examination for Clinical Medicine and inputted them into ChatGPT, which simulates scenarios where a student answers a common medical question without any choices or a doctor tries to diagnose a patient based on multimodal clinical data (ie, symptoms, history, physical examination, and laboratory values). ChatGPT’s accuracy for all questions was 31.5%. Among the 3 subgroups, namely, CQs, MCQs, and CAQs, the accuracy was 42%, 37%, and 17%, respectively (Figure 2). Compared to the original questions, the accuracy of the encoding questions decreased by 19%, 17%, and 14% for CQs, MCQs, and CAQs, respectively, which demonstrates that the ability of ChatGPT to answer the open-ended questions in Chinese is a shortcoming. During the adjudication stage, there was substantial agreement among physicians on prompts in all 3 subgroups (κ ranged from 0.80 to 1.00).

ChatGPT Demonstrates High Internal Concordance

Concordance, which is a measure of the level of agreement or similarity between the option selected by AI and its subsequent explanation, was also taken into consideration. The results showed that ChatGPT achieved 90% concordance across all questions, and this high concordance was maintained across all 3 subgroups (Figure 3). Additionally, we analyzed the concordance difference between correct and incorrect answers and found that concordance among accurate responses was perfect and significantly greater than among inaccurate responses (n=52, 100% vs n=113, 50%; P<.001; Figure 3). These findings suggest that ChatGPT has a high level of answer-explanation concordance in Chinese, likely due to its strong internal consistency in its probabilistic language model.
Figure 3. Concordance of ChatGPT on Chinese Postgraduate Examination for Clinical Medicine after encoding. For the subgroup “case analysis question,” artificial intelligence outputs were adjudicated to be concordant and discordant based on the scoring system provided in Table S2 in Multimedia Appendix 1 data. It demonstrates concordance rates stratified between accurate, inaccurate, and indeterminate outputs across all the case analysis questions.

ChatGPT Shows Multiple Insights Toward the Same Questions
Another evaluation index considered was the frequency of insights generated by the AI model that quantifies the quantity of insights produced. After evaluating the score, accuracy, and concordance of ChatGPT, its potential was investigated to enhance medical education by augmenting human learning. We analyzed the frequency of insights provided by ChatGPT. Remarkably, ChatGPT generated at least 1 significant insight in 80% (n=132) of all questions (Figure 4). Moreover, the analysis revealed that the accuracy response had the highest frequency of insights with an average of 2.95. The indeterminate response followed closely behind with an average of 2.7, while the inaccurate response had a lower frequency of insights with an average of 1.39 (Figure 4). The high frequency of insights in the accurate group suggests that it may be feasible for a target learner to acquire new or remedial knowledge from the ChatGPT AI output.

Figure 4. The frequency of insights of ChatGPT on Chinese Postgraduate Examination for Clinical Medicine after encoding. For the subgroup "case analysis question," artificial intelligence outputs were adjudicated to count the frequency of insights it offered. It demonstrates the frequency of insights stratified between accurate, inaccurate, and indeterminate outputs, across all the case analysis questions.

Discussion
Major Findings
To evaluate ChatGPT’s problem-solving capabilities and assess its potential for Chinese medical education integration, its performance on the Chinese Postgraduate Examination for Clinical Medicine was tested. We had two major findings: (1) the score of ChatGPT needs to be improved when facing questions asked in the Chinese language and (2) there is still potential for this AI to generate novel performance that can assist humans due to the high concordance and the frequency of insights. This is the first study to assess the performance of ChatGPT in medical care and clinical decisions in Chinese.
ChatGPT’s Performance Needs Improvement for Medical Questions in Chinese

A recent study showed that ChatGPT (version 3.5) performed with an accuracy rate of over 50% across all examinations and even exceeded 60% accuracy in some analyses when facing the USMLE [7]. Our results indicate that ChatGPT exhibited moderate accuracy in answering open-ended medical questions in Chinese, with an accuracy of 31.5%. Given the differences between English and Chinese inputs, it suggests that ChatGPT requires further improvement in answering medical questions in the Chinese language.

We sought to understand why there is a significant discrepancy between the performance of ChatGPT on Chinese and English language examinations. To investigate this, we asked the ChatGPT for the reasons, it explains that the training data used to train AI in different languages may be different, and the algorithms used to process and analyze text may vary from language to language (data not shown). Therefore, even for the same question, the output generated may vary slightly based on the language and the available language-based data.

Upon analyzing the results of this research, we found that the accuracy of ChatGPT was lowest for MCQs, followed by CQs and CAQs. The lower accuracy on MCQs may be due to the model being undertrained on the input as well as the MCQ samples being significantly less than those of single-choice questions. On the other hand, the CAQs may have extensive training compared to MCQs and are similar in type to the USMLE question.

Furthermore, we noticed that high accuracy outputs were associated with high concordance and a high frequency of insight, whereas poorer accuracy was linked to lower concordance and a lack of insight. Thus, it was hypothesized that inaccurate responses were primarily driven by missing information, which could result in reduced insight and indecision in the AI, rather than an overcommitment to an incorrect answer [7]. The results indicate that enhancing the database and providing additional training with Chinese questions could substantially improve the performance of the model.

Challenges of AI in Future Applications

Despite the promising potential of AI in medicine, it also poses some challenges. Standards for using AI in health care still need to be developed [8,9], including clinical care, quality, safety, malpractice, and communication guidelines. Furthermore, the implementation of AI in health care requires a shift in medical culture, which poses a challenge for both medical education and practice. Additionally, ethical considerations must be taken into account, such as data privacy, informed consent, and bias prevention, to ensure that AI is used ethically and for the benefit of patients. Surprisingly, a recently launched AI system for autonomous detection of diabetic retinopathy carries medical malpractice and liability insurance [10].

Prospective of AI

AI is a rapidly growing technology. At the time of writing, ChatGPT (version 4) has been released with significant improvements. Numerous practical and observational studies have demonstrated the versatile role of AI in almost all medical disciplines and specialties, particularly in improving risk assessment [11,12], data reduction, clinical decision support [13,14], operational efficiency, and patient communication [15,16]. We anticipate that advanced language models such as ChatGPT are reaching a level of maturity that will soon have a significant impact on clinical medicine, enhancing the delivery of personalized, compassionate, and scalable health care.

A comparison of ChatGPT’s performance with other AI models, particularly in the context of Chinese language performance, could yield more comprehensive insights and underscore the unique challenges of using AI in diverse linguistic landscapes. However, this was primarily due to the fact that AI models that focus on other aspects, while enhancing medical education and achieving promising results in medical question answering, are mostly developed and evaluated using English language data sets. This limitation restricts their applicability for performance comparisons in the context of the Chinese language.

Limitations

One limitation of this research is the small sample size. We only accessed 165 samples to qualify its accuracy and 30 CAQs to qualify its concordance and frequency of insight due to the limitations of the data, which focused solely on the diagnosis of the patient. Furthermore, the clinical situation is more complicated than the test, and larger and deeper analyses were needed. Finally, bias and error were inevitable in human adjudication, although there was a good interrater agreement between the physicians for the adjudication.

Moreover, comparing ChatGPT’s performance with other AI models, especially in the context of Chinese language, can provide valuable insights and highlight the distinctive challenges associated with leveraging AI in diverse linguistic environments. One notable factor contributing to this need for comparison is the prevalence of AI models such as Bidirectional Encoder Representations from Transformers, CLUE-Med, and MedQA that have made significant contributions to medical education and demonstrated promising outcomes in medical question answering. However, these models have predominantly been developed and assessed using English language data sets. This particular limitation hampers their suitability for conducting performance assessments within the Chinese language domain.

Conclusions

In conclusion, although the ChatGPTs got a score over the passing score in the Chinese Postgraduate Examination for Clinical Medicine, the performance was limited when presented with open-ended questions. On the other hand, ChatGPT demonstrated a high level of internal concordance, which suggests that the explanations provided by ChatGPT support and affirm the given answers. Moreover, ChatGPT generated multiple insights toward the same questions, demonstrating its potential for generating a variety of useful information. Further prospective studies are needed to explore whether there is a language-based difference in the performance of medical education settings and clinical decision-making, such as Chinese and minority languages.
Acknowledgments
The authors acknowledge ChatGPT for polishing their paper.

Data Availability
All data generated or analyzed during this study are included in this published paper (and Multimedia Appendix 1).

Authors’ Contributions
Xiao Liu was responsible for the entire project and revised the draft. CF, AA, and YW performed the data extraction, statistical analysis, and interpretation of the data. WZ, Z Chen, YZ, and JW drafted the first version of the paper. All authors participated in the interpretation of the results and prepared the final version of the paper.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Original questions, adjudication criteria for accuracy and concordance, and \( \kappa \) statistic for interrater agreement between adjudicating physicians.

References


Abbreviations

AI: artificial intelligence
CAQ: case analysis question
CQ: common question
MCQ: multiple-choice question
USMLE: United States Medical Licensing Examination

©Peng Yu, Changchang Fang, Xiaolin Liu, Wanying Fu, Jitao Ling, Zhiwei Yan, Yuan Jiang, Zhengyu Cao, Maoxiong Wu, Zhiteng Chen, Wengen Zhu, Yuling Zhang, Ayiguli Abudukeremu, Yue Wang, Xiao Liu, Jingfeng Wang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Abstract

Background: The use of mobile devices for delivering health-related services (mobile health [mHealth]) has rapidly increased, leading to a demand for summarizing the state of the art and practice through systematic reviews. However, the systematic review process is a resource-intensive and time-consuming process. Generative artificial intelligence (AI) has emerged as a potential solution to automate tedious tasks.

Objective: This study aimed to explore the feasibility of using generative AI tools to automate time-consuming and resource-intensive tasks in a systematic review process and assess the scope and limitations of using such tools.

Methods: We used the design science research methodology. The solution proposed is to use cocreation with a generative AI, such as ChatGPT, to produce software code that automates the process of conducting systematic reviews.

Results: A triggering prompt was generated, and assistance from the generative AI was used to guide the steps toward developing, executing, and debugging a Python script. Errors in code were solved through conversational exchange with ChatGPT, and a tentative script was created. The code pulled the mHealth solutions from the Google Play Store and searched their descriptions for keywords that hinted toward evidence base. The results were exported to a CSV file, which was compared to the initial outputs of other similar systematic review processes.

Conclusions: This study demonstrates the potential of using generative AI to automate the time-consuming process of conducting systematic reviews of mHealth apps. This approach could be particularly useful for researchers with limited coding skills. However, the study has limitations related to the design science research methodology, subjectivity bias, and the quality of the search results used to train the language model.

Introduction

The delivery of health-related services through the use of mobile devices (mHealth) [1] has been growing at a tremendous pace.
conducted a preliminary search on the US federal clinical trials database (ClinicalTrials.gov) and had to combine the keywords “mHealth,” “mobile,” and “cell AND phone” to obtain 1678 studies and their results. Today, that same number can be obtained using “mHealth” alone as a keyword. As the need for mHealth evidence has grown, so too has the necessity for summarizing both the state of the art and the practice.

Systematic reviews seek to collect and combine relevant evidence within the specific scope of a research question while also striving to minimize bias [3,4]. In PubMed alone, the number of systematic reviews published on digital health–related topics has increased a hundredfold in the last 10 years. In fact, the pace at which the mHealth field is developing for certain conditions like breast cancer is such that systematic reviews can be found every 2 or 3 years [5-9]. The systematic review process, however, is a time- and resource-intensive process, reportedly requiring a median of 5 researchers and approximately 40 weeks of work to reach submission [10-12].

The emergence of generative AI has been seen as a breakthrough in the field of automation. With the ability to generate content such as text, images, and even music, AI has been reported as a potential solution to tedious time-consuming and labor-intensive tasks [13]. For instance, generative AI can be used to automatically generate product descriptions, news articles, or even code [14]. By eliminating the need for human intervention, generative AI can free up valuable time and resources for more complex tasks, thereby improving efficiency and accuracy. ChatGPT, a natural language processing model with a capacity of 175 billion parameters, has been trained on extensive amounts of data and is designed to produce human-like responses to user inputs. Since its release in November 2022, ChatGPT has received significant attention from media and academia alike, provoking ethical discussions on scientific authorship [15,16], attempting to pass medical license and specialist examinations [17-19], and even designing medical education curricula [20].

The objective of this study was to explore the feasibility of using generative AI tools to automate time-consuming and resource-intensive tasks in a systematic review process and assess the scope and limitations of using such tools.

**Methods**

**Study Design**

This study uses a design science research (DSR) methodology. DSR is a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts [21]. DSR commonly involves the identification of a problem or opportunity, followed by the development, implementation, and evaluation of a solution. In DSR, as well as in action research, the process happens within an organization that provides context and that would be changed as a result of the use of the artifact [21]. An overview of the process adapted from Hevner [22] can be seen in Figure 1.

![Figure 1. Design science research overview, adapted from Hevner 2004 [22].](https://mededu.jmir.org/2024/1/e48949)

**Problem Definition**

The problem to which DSR was applied was the time-consuming and resource-intensive process of conducting systematic reviews of mHealth applications.

**Organizational Context**

The organizational context consisted of the More Stamina team of researchers, software developers, and health care professionals, working collaboratively within the host research institutions (ie, the University of Oulu and Trinity College Dublin).

The More Stamina project aims to create an evidence-driven gamified mHealth solution for people with multiple sclerosis (MS), where each step of the development follows a scientific process, as follows: MS needs as well as barriers and facilitators were explored through qualitative studies [23]; the state of the practice for MS apps was systematically reviewed [24,25]; user-centered design was used to create “MS personas” [23];
Cocreation sessions took place to produce solution concepts [26]; the design, prototyping, and initial usability testing were described [27]; early health technology assessment was used to guide software development [28]; patient representatives were involved throughout the project [29]; and user testing and feasibility studies were ongoing in a multicenter study [30].

A script using the software application for audience targeting called 42matters [31] was used in the past to extract information from different app stores. The script is no longer functional, and person-hours from the software development team were not able to be dedicated to this task.

**Background Studies**

The research plans and outcomes from previous studies, where systematic review methodologies were used to identify, select, collect, and analyze features and content of mHealth apps [6,24,25], served as models for our study. In those studies, a search strategy was defined, using relevant main keywords for each condition. App stores were searched, taking steps to ensure that no previous search history or cookies influenced results. Screening took place based on mHealth applications’ titles, descriptions, and metadata.

### Table 1. Digital skills background.

<table>
<thead>
<tr>
<th>Competency</th>
<th>Level</th>
<th>Experience</th>
<th>Self-assessment score (of 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scrum master</td>
<td>Certified Scrum Master</td>
<td>Agile methodologies and team management</td>
<td>7</td>
</tr>
<tr>
<td>Product owner</td>
<td>Certified Scrum Product Owner</td>
<td>Product road mapping and stakeholder management</td>
<td>8</td>
</tr>
<tr>
<td>Game design</td>
<td>Intermediate</td>
<td>Game mechanics, storytelling, and level design</td>
<td>7</td>
</tr>
<tr>
<td>Web design</td>
<td>Advanced</td>
<td>User experience and user interface design and responsive design</td>
<td>8</td>
</tr>
<tr>
<td>JavaScript</td>
<td>Beginner</td>
<td>Front-end development</td>
<td>5</td>
</tr>
<tr>
<td>HTML5</td>
<td>Intermediate</td>
<td>Front-end development</td>
<td>7</td>
</tr>
<tr>
<td>PHP: Hypertext Preprocessor</td>
<td>Beginner</td>
<td>Front-end and back-end development</td>
<td>4</td>
</tr>
</tbody>
</table>

**Evaluation**

The results were evaluated for their effectiveness and efficiency in replicating the initial steps of the background studies. The measurements were considered with respect to the amount of time required to generate a spreadsheet containing the necessary information for human reviewers to start the systematic review process. The output was compared to the output generated by the previous script, which required further processing of the data.

**Ethical Considerations**

No ethics board review is needed as the work does not use patient data or involve human participants.

**Results**

The following section reports the overall design, development, and implementation process as well as the evaluation results and highlights of the experience. Prompts provided by GG and generative AI are tagged accordingly.
Following the premise of the cocreation goal, ChatGPT was asked (by GG) “I am new to Python, how do I run a Python script?” to which the response was the following:

To run a Python script, you need to follow these steps:
1. Install Python on your computer, if it’s not already installed. You can download Python from the official website: https://www.python.org/downloads/.
2. Open a text editor (e.g. Notepad, Sublime Text, Atom) and create a new file.
3. Write your Python code in the file.
4. Save the file with a .py extension (e.g. “my_script.py”).
5. Open a command prompt or terminal and navigate to the directory where your Python script is saved.
6. Type “python” followed by the name of your script and press enter. For example, if your script is named “my_script.py”, you would type “python my_script.py” and press enter.
7. Your Python script will now run, and you should see any output or results in the command prompt or terminal.

Note: If you’re using an Integrated Development Environment (IDE) like PyCharm or Visual Studio Code, you can run your script directly from the IDE by clicking on the “Run” button or pressing a keyboard shortcut. [ChatGPT]

ChatGPT provided step-by-step guidance for the installation and setup of Visual Studio Code, along with the required libraries and dependencies. After the development environment was ready, the suggested Python script was copied and pasted to check its functionality. The code produced several errors, which were fed back to ChatGPT by GG, as follows:

I am getting “ModuleNotFoundError: No module named ‘requests’”
Now it says “ModuleNotFoundError: No module named ‘bs4’”

ChatGPT stated that the “requests” library was imported but not used, so it should be removed. Subsequent error messages were troubleshooted with the AI. At times, the script would stop responding due to a time-out error or issues with the output. To make the debugging process easier to follow and understand, ChatGPT was prompted by GG to do the following:

add error messages that explain what went wrong.
add indicators for each step that it is working.  

could you add a progress bar indicator to each step?  

I would like the responses to be colored.

Upon review, it was clear that some error caused the output files to be overwritten. The author requested the following: “how would you add a condition to check whether a file with the same name already exists and if it does ask whether to overwrite it?” The AI added the pertinent libraries to the script and produced an input prompt for the user (Figure 3).

Figure 3. Added libraries and code for checking existing files.

The triggering prompt expected the script to not only be searching app stores for matches with “multiple sclerosis” but also to cross-reference them with scientific publications found in PubMed. As described in the Background Studies section of this paper, this process was done by looking for keywords present in the app description. As it were, the script was designed to only search for “multiple sclerosis” and crashed when it reached the PubMed phase. ChatGPT diagnosed a time-out error due to the number of publications and provided a potential solution to improve the indicator bars:

The progress bar for PubMed search may not be moving because the total value for it is set to 3600, which is the number of seconds in an hour. This is the unit being used for the progress bar, not the total number of items to be searched. To fix this, you should update the total value for the progress bar to be the total number of apps being searched, which is the length of the app_df dataframe multiplied by the maximum number of PubMed IDs to be retrieved for each app (100 in this case). You can update the progress bar as follows... [ChatGPT]

After these issues were sorted and the script could properly fetch PubMed results, more keywords were entered into the script by prompting ChatGPT with the following:

how would you make it so that the items in the medical_keywords list are taken from a csv file called “keywords.csv”? [GG]

A screenshot of the final Python script running can be seen in Figure 4.
Evaluation

As explained in the Background Studies section of this paper, app data extraction from the Google Play Store resulted in a spreadsheet file that contained the mHealth app’s name, store link, app description, developer’s name, developers’ URL, price, number of downloads, and app rating. During the screening phase of the studies, the research team read the apps’ descriptions and flagged those that contained keywords or sentences suggestive of the evidence base for in-depth review.

The ChatGPT-generated code resulted in a CSV file that contained the app’s name, store link, app description, and a column titled “Medical Evidence.” There were no columns containing metadata, and the Medical Evidence column only contained “Yes” or “No,” accordingly. Closer inspection revealed that the script was searching for a full match on the apps’ titles in PubMed results. The resulting document was useful as an intermediate outcome but was deemed unsuitable as a final output. The overall cocreation process had a total duration of 4 hours and 39 minutes, providing a working script version available on GitHub [32].

Using the results from the ChatGPT-generated script to fully automate the process would likely require further work refining the script, either by using the steps of the background studies to base the script or by providing clearer starting prompts for the generative AI. However, leveraging this approach as a means to advance work when the software developing team was otherwise engaged was useful.

Highlights

Some highlights of this study are as follows:

- The overall cocreation process exercise had a total duration of 4 hours and 39 minutes.
- There were several misunderstandings during the interactions, not unlike the challenges one might encounter when messaging a more experienced coder.
- Structured thinking ahead of time reduced the number of misunderstandings.
- No knowledge of Python scripting was required by the author.
- The resulting output was useful to continue a systematic review but not sufficient to replace the final outputs.

Discussion

Principal Results

This study is the first to describe the cocreation process with a generative AI in developing an automated script for conducting a systematic review of mHealth apps. The study provides insights into the potential of using this kind of AI tools for researchers with little to no coding skills, and it identifies an innovative way of approaching a research problem and facilitating interdisciplinary collaborations. This study also makes a methodological contribution, expanding knowledge as it uses DSR, an approach that is not commonly used in health care and health informatics [33].

Comparison With Prior Work

The resource-intensive process and the burden that systematic reviews represent have been highlighted in the literature before. The use of multiple databases, such as MEDLINE, Embase, Cochrane Library, and Web of Science as well as clinical trial registries like ClinicalTrials.gov are common practices to increase results [34]. However, this tactic requires a lengthy deduplication process, involving long manual procedures, potentially introducing quality-affecting errors and biases [35-37]. In fact, automation attempts using AI models have been made in the past, with a focus on the deduplication problem, as seen in studies by Borissov et al [38] and Bramer et al [39].

Performing a systematic review is a common step in doctoral researchers’ studies [40,41], as a means of introducing the candidate to the topic. The use of generative AI to cocreate scripts like the one presented in this study could help automate the time-consuming process, allowing researchers to focus on other aspects of the research process.

The ethical implications of using generative AI models, such as ChatGPT, to generate scientific authorship have sparked discussions [15,16]. AI’s potential for assisting in academic research needs to be considered and weighed against the potential for its misuse. Although generative AI can assist in the development of a systematic review script, it is important to note that the final review still requires human oversight and input to not only assess the accuracy and relevance of the results but also ensure that the ethical principles have been followed.
Beyond research, there are wider implications for the use of generative AI in both medical education and the upskilling of the health care workforce. The need for more digital skills training for health care professionals is widely recognized [42], and other authors have further explored medical degree programs’ curricula to examine how AI is included [43,44]. A recent publication explored the specific competencies needed for the effective and ethical use of AI in health care [45]. Understanding basic knowledge of AI and its applications as well as how to integrate AI into the general workflow of different tasks ranked among the top 6 key competency domains.

The role of generative AI in evolving health care education is pivotal, especially as universities adapt to its challenges. Generative AI has the potential to streamline processes like systematic reviews and clinical information retrieval, thereby allowing health care professionals to focus more on patient-centered, empathetic care and the co-design of effective treatment outcomes.

**Limitations**

The results of this study must be considered within its limitations. The DSR methodology was developed for this specific problem, which limits applicability in other contexts.

In addition, subjectivity is a common bias present in DSR, which can make it difficult to establish the reliability and validity of the results. The main goal of DSR is to generate prescriptive knowledge, which provides guidelines on how to effectively design and implement solutions in the organizational context. However, as DSR focuses more on developing practical solutions rather than generating new theoretical insights, it was aligned with the goal of this study. DSR differs from traditional research paradigms by focusing more on creating and evaluating new solutions rather than on understanding existing phenomena. Further, while generative AI can assist in the development of a systematic review script, the result will be greatly affected by the training data used for the language model. Additionally, there may be limitations in the quality of the search results obtained from the previous studies, which only become apparent through automated processes.

**Conclusions**

This study outlined the cocreation process of an automated script for systematic reviews of mHealth apps, using generative AI. The study shed light on the potential of such AI tools for researchers with limited coding abilities and highlighted a novel approach for addressing research problems and promoting interdisciplinary collaborations.

**Acknowledgments**

GG would like to thank Prof Octavio Rivera-Romero, Dr Estefania Guisado-Fernandez, Dr Diego Giunta, Dr Analia Baum, and Prof Minna Isomursu for their collaboration and support.

This study has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie grant agreement (101034252). The publication has also emanated from research supported (in part) by a research grant from Science Foundation Ireland (SFI) under grant number 16/RC/3948 and Business Finland’s More Stamina Research to Business project.

The authors are grateful for ChatGPT, whose collaboration was essential for the completion and inception of this study.

**Conflicts of Interest**

None declared.

**References**


Abbreviations
AI: artificial intelligence
DSR: design science research
mHealth: mobile health
MS: multiple sclerosis

©Guido Giunti, Colin P Doherty. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,
Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models

Tassallah Abdullahi¹, MSc; Ritambhara Singh¹,², PhD; Carsten Eickhoff³, PhD

¹Department of Computer Science, Brown University, Providence, RI, United States
²Center for Computational Molecular Biology, Brown University, Providence, RI, United States
³School of Medicine, University of Tübingen, Tübingen, Germany

Corresponding Author:
Carsten Eickhoff, PhD
School of Medicine
University of Tübingen
Schaffhausenstr, 77
Tübingen, 72072
Germany
Phone: 49 7071 29 843
Email: carsten.eickhoff@uni-tuebingen.de

Abstract

Background: Patients with rare and complex diseases often experience delayed diagnoses and misdiagnoses because comprehensive knowledge about these diseases is limited to only a few medical experts. In this context, large language models (LLMs) have emerged as powerful knowledge aggregation tools with applications in clinical decision support and education domains.

Objective: This study aims to explore the potential of 3 popular LLMs, namely Bard (Google LLC), ChatGPT-3.5 (OpenAI), and GPT-4 (OpenAI), in medical education to enhance the diagnosis of rare and complex diseases while investigating the impact of prompt engineering on their performance.

Methods: We conducted experiments on publicly available complex and rare cases to achieve these objectives. We implemented various prompt strategies to evaluate the performance of these models using both open-ended and multiple-choice prompts. In addition, we used a majority voting strategy to leverage diverse reasoning paths within language models, aiming to enhance their reliability. Furthermore, we compared their performance with the performance of human respondents and MedAlpaca, a generative LLM specifically designed for medical tasks.

Results: Notably, all LLMs outperformed the average human consensus and MedAlpaca, with a minimum margin of 5% and 13%, respectively, across all 30 cases from the diagnostic case challenge collection. On the frequently misdiagnosed cases category, Bard tied with MedAlpaca but surpassed the human average consensus by 14%, whereas GPT-4 and ChatGPT-3.5 outperformed MedAlpaca and the human respondents on the moderately often misdiagnosed cases category with minimum accuracy scores of 28% and 11%, respectively. The majority voting strategy, particularly with GPT-4, demonstrated the highest overall score across all cases from the diagnostic complex case collection, surpassing that of other LLMs. On the Medical Information Mart for Intensive Care-III data sets, Bard and GPT-4 achieved the highest diagnostic accuracy scores, with multiple-choice prompts scoring 93%, whereas ChatGPT-3.5 and MedAlpaca scored 73% and 47%, respectively. Furthermore, our results demonstrate that there is no one-size-fits-all prompting approach for improving the performance of LLMs and that a single strategy does not universally apply to all LLMs.

Conclusions: Our findings shed light on the diagnostic capabilities of LLMs and the challenges associated with identifying an optimal prompting strategy that aligns with each language model’s characteristics and specific task requirements. The significance of prompt engineering is highlighted, providing valuable insights for researchers and practitioners who use these language models for medical training. Furthermore, this study represents a crucial step toward understanding how LLMs can enhance diagnostic reasoning in rare and complex medical cases, paving the way for developing effective educational tools and accurate diagnostic aids to improve patient care and outcomes.

https://mededu.jmir.org/2024/1/e51391
Introduction

Background

Natural language processing has witnessed remarkable advances with the introduction of generative large language models (LLMs). In November 2022, OpenAI released ChatGPT-3.5 (OpenAI), a large natural language processing chatbot trained on a large corpus collected from the internet to generate humanlike text in response to user queries. ChatGPT-3.5 has seen massive popularity, and users have praised its creativity and language comprehension for several tasks, such as text summarization and writing computer programs [1]. In March 2023, OpenAI responded to the success of ChatGPT-3.5 by introducing an enhanced iteration called GPT-4, specifically designed to address intricate queries and nuanced directives more effectively. Shortly thereafter, Google released their comparable model, Bard (Google LLC), which joined the league of impressive LLMs. What sets Bard apart is its real-time access to and use of internet information, enriching its response generation with up-to-date information [2]. In contrast, GPT-4 possesses multimodal capabilities, including image inputs, albeit not publicly available during the study [3].

These LLMs were not originally designed for medical applications. However, several studies [4-5] have shown their extraordinary capabilities in excelling in various medical examinations, such as the Self-Assessment in Neurological Surgery examination and the USMLE (United States Medical Licensing Examination). Their results demonstrated the ability of these models to handle clinical information and complex counterfactuals. Furthermore, numerous investigations [6-8] have revealed the remarkable advantages of harnessing the power of LLMs in diverse medical scenarios. Notably, Lee et al [8] demonstrated using LLMs as a reliable conversational agent to collect patient information to assist in medical notetaking, whereas Patel and Lam [9] delved into using LLMs as a valuable tool for generating comprehensive patient discharge summaries. The ability of LLMs to process and generate medical text has unlocked new opportunities to enhance diagnostic reasoning, particularly in tackling rare and complex medical cases.

Rare diseases are characterized by their low prevalence in the general population, whereas complex diseases are conditions with overlapping factors and multiple comorbidities that are often difficult to diagnose [10,11]. Sometimes, a condition can be rare and complex if it is infrequent and challenging to diagnose accurately [11]. Rare and complex diagnoses present significant challenges across various medical levels and often require extensive medical knowledge or expertise for accurate diagnosis and management [10,11]. This may be because, during their education, physicians are trained to prioritize ruling out common diagnoses before considering rare ones during patient evaluation [12]. In addition, most medical education programs rarely cover some complex conditions, and guidance for practicing clinicians is often outdated and inappropriate [13,14]. As a result, most physicians perceive their knowledge of rare diseases as insufficient or very poor, and only a few feel adequately prepared to care for patients with these conditions [12,15]. This knowledge gap increases the risk of misdiagnosis among individuals with rare and complex conditions. Furthermore, the scarcity of available data and the relatively small number of affected individuals create a complicated diagnostic landscape, even for experienced and specialized clinicians [10]. Consequently, patients often endure a prolonged and arduous diagnostic process. Therefore, there is a pressing need for comprehensive educational tools and accurate diagnostic aids to fill the knowledge gap and address these challenges effectively.

This study aims to explore the potential of 3 LLMs, namely Bard, GPT-4, and ChatGPT-3.5, as continuing medical education (CME) systems to enhance the diagnoses of rare and complex conditions. Although these models have demonstrated impressive success in standardized medical examinations [4,5], it is important to acknowledge that most examinations reflect general clinical situations, which may not fully capture the intricacies encountered in real-world diagnostic scenarios. Furthermore, these standardized tests often feature questions that can be answered through memorization [16]. In contrast, real-world complex diagnostic scenarios that physicians face involve dynamic, multifaceted patient cases with numerous variables and uncertainties. Although previous studies by Liu et al [17] and Cascella et al [18] have highlighted the ability of LLMs to support health care professionals in real-world scenarios, their effectiveness in diagnosing rare and complex conditions remains an area of exploration. Despite the promising use of LLMs in medical applications, studies have reported that their responses to user queries are often nondeterministic (ie, depending on the query format) and exhibit significant variance [17,19]. This attribute may pose challenges in clinical decision support scenarios because the dependability of a system is uncertain when its behavior cannot be accurately predicted. However, no investigation has been conducted to show how different input formats (prompts) affect LLM responses in the medical context.

Prompt engineering is a technique for carefully designing queries (inputs) to improve the performance of generative language models [20,21]. We can guide LLMs to generate more accurate and reliable responses by carefully crafting effective prompts. Our study investigated effective prompting strategies to improve the accuracy and reliability of LLMs in diagnosing rare and complex conditions within an educational context. We evaluated the performance of LLMs by comparing their responses to those of human respondents and the responses of
MedAlpaca [22], an open-source generative LLM designed for medical tasks. Given the documented advantages of using LLMs as a complementary tool rather than a substitute for clinicians [17,18], our study incorporated LLMs with the understanding that clinicians may use them beyond real-time diagnostic scenarios. Although our premise is based on a clinician having established an initial diagnostic hypothesis and seeking further assistance to refine the precise diagnosis, we acknowledge the broader utility of LLMs. They can be valuable in real-time decision support and retrospective use during leisure or documentation, allowing physicians to experiment with and enhance their understanding of rare and complex diseases. This approach recognizes the inherent uncertainty in diagnosis and harnesses the capabilities of LLMs to assist clinicians in various aspects of their diagnostic processes. In the context of CME, our study highlights the possibility of integrating LLMs as a valuable addition. By providing further assistance in refining complex and rare diagnoses, these LLMs could support evidence-based decision-making among health care professionals for improved patient outcomes.

**Objectives**

Our study has 2 main objectives: first, to examine the potential of LLMs as a CME tool for diagnosing rare and complex conditions, and second, to highlight the impact of prompt formatting on the performance of LLMs. Understanding these aspects could significantly contribute to advancing diagnostic practices and effectively using LLMs to improve patient care.

**Methods**

**Data Sets**

We used 2 data sets to examine the capacity of LLMs to diagnose rare and complex conditions as follows:

1. Diagnostic case challenge collection (DC3) [11] comprises 30 complex diagnostic cases curated by medical experts in the *New England Journal of Medicine* web-based case challenges. The original cases contained text and image descriptions of patients’ medical history, diagnostic imaging, and laboratory results; however, we used only textual information to form prompts (queries). The web-based polls recorded an average of 5850 (SD 2522.84) respondents per case, many of whom were health care professionals. The participants were required to identify the correct diagnosis from a list of differential diagnoses. Case difficulty was categorized based on the percentage of correct responses received from the respondents on the web-based survey. The case categories were: “rarely misdiagnosed cases” (with ≥21/30, 70% correct responses), “moderately misdiagnosed cases” (with >9/30, 30% and <21/30, 70% correct responses), and “frequently misdiagnosed cases” (with ≤9/30, 30% correct responses). Furthermore, the final diagnoses determined by the treating physicians of the cases were provided alongside the poll results, enabling the comparison of the performance of human respondents with that of the targeted LLMs.

2. Medical Information Mart for Intensive Care-III (MIMIC-III) [23] comprises deidentified electronic health record data from approximately 50,000 Boston Beth Israel Deaconess Medical Center intensive care unit patients. We focused on discharge summaries containing the accumulated patient information from admission to discharge. Similar to previous work on clinical outcome prediction by van Aken et al [24] and Abdullahi et al [25], we filtered document sections unrelated to admissions, such as discharge information or hospital course and retained sections related to admissions, such as chief complaint, history of illness or present illness, medical history, admission medications, allergies, physical examination, family history, and social history. Each discharge summary had a discharge diagnosis section that indicated the patient’s final diagnosis for that admission. We reviewed the discharge summaries to identify rare diseases and referred to the Orphanet website [26]. In this study, we randomly selected 15 unique, rare conditions as our target. These cases were selected as pilot studies for a focused and in-depth analysis.

**Models**

In this study, we conducted experiments using LLMs designed for conversational context. Specifically, we used the July 6, 2023, version of Bard; the July 4, 2023, versions of GPT-4 and ChatGPT-3.5; and the publicly available version of MedAlpaca 7b [22]. We entered prompts individually through the chat interface to evaluate Bard, GPT-4, and ChatGPT-3.5, treating each prompt as a distinct conversation. MedAlpaca differs from Bard, ChatGPT-3.5, and GPT-4 in that it requires users to submit queries or prompts through a Python (Python Software Foundation) script. Consequently, we used a single Python script for each prompt strategy to submit queries for each data set. It is worth noting that Bard has certain limitations compared with ChatGPT-3.5 and GPT-4. Bard has a restricted capacity to handle lengthy queries. Moreover, Bard is more sensitive to noisy input and specific characters. For example, the MIMIC-III data set contained deidentified patients’ notes filled with special characters such as “[**Hospital 18654**]” and laboratory results written in shorthand, for example, * Hgb-9.6* Hct-29.7* MCV-77* MCH-24.9*. Consequently, to work effectively with Bard, we preprocessed the text by removing special characters and retaining only alphanumeric characters.

**Prompting Strategies**

Direct (standard prompting) and iterative prompting (chain of thought prompting) [27] are the 2 major prompting methods. Iterative prompting is a promising method for improving LLM performance on specialized tasks; however, it requires a predefined set of manually annotated reasoning steps, which can be time consuming and difficult to create, especially for specialized domains. Most users opt for a direct prompt method to save time and obtain an immediate response. Therefore, to analyze the effect of prompt formats on LLM performance, we assessed each model’s performance for every case using the 3 distinct direct prompt strategies outlined in Table 1. These strategies varied from open-ended to multiple-choice formats.
Table 1. Prompt strategies.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Prompt strategy description</th>
<th>Prompt sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach 1 (open-ended prompt)</td>
<td>In this approach, prompts were formatted in an open-ended fashion. Formatting a prompt using this method allows the model to formulate a hypothesis for the case and explain why and what it thinks is the diagnosis. Here, we scored a model based on its ability to provide the correct diagnosis without additional assistance.</td>
<td>“What is the diagnosis? The case is: A 32-year-old man was evaluated in the emergency department of this hospital for the abrupt onset of postprandial chest pain...”</td>
</tr>
<tr>
<td>Approach 2 (multiple-choice prompt)</td>
<td>We formatted prompts as multiple-choice questions, and the LLMs were expected to select a single diagnosis from a list of options. The models were assigned a positive score in this task if they selected the correct diagnosis from the options.</td>
<td>“Choose the most likely diagnosis from the following: Option I: Cholecystitis, Option II: Acute coronary syndrome, Option III: Pericarditis, Option IV: Budd-Chiari syndrome. The case is: A 32-year-old man was evaluated in the emergency department of this hospital for the abrupt onset of postprandial chest pain...”</td>
</tr>
<tr>
<td>Approach 3 (ranking prompt)</td>
<td>The prompts were presented as a case and a list of diagnoses to be ranked by the LLMs. Models were assigned a positive score if the correct diagnosis was ranked first in this format.</td>
<td>“Rank the following diagnoses according to the most likely. Option I: Cholecystitis, Option II: Acute coronary syndrome, Option III: Pericarditis, Option IV: Budd-Chiari syndrome. The case is: A 32-year-old man was evaluated in the emergency department of this hospital for the abrupt onset of postprandial chest pain...”</td>
</tr>
</tbody>
</table>

aLLM: large language model.

Building upon prior research by Wang et al [28] and Li et al [29], we hypothesized that using a diverse range of prompts can reveal distinct reasoning paths while maintaining consistency in the correct responses regardless of the variations. When using multiple-choice prompts for the DC3 cases, we presented the same options available in the original web-based polls to the models, but on the MIMIC-III data set, we generated random wrong answers that were closely related to the correct diagnosis. We evaluated each LLM by assigning a positive or negative score (binary score) based on their responses. A positive score was assigned only if the models correctly selected the diagnosis for either data set. Conversely, we omitted the options for open-ended prompts, expecting the models to generate the correct diagnosis independently. Positive scores were awarded only if the models accurately provided the correct diagnosis.

**Prompt Ensemble: Majority Voting**

To safely use imperfect language models, users must determine when to trust their predictions, particularly in critical situations, such as clinical decision support. Therefore, we used a majority voting (prompt ensembling) strategy to enhance the reliability of LLMs’ responses. The majority voting approach involves aggregating multiple responses and selecting the most common answer. By applying this approach to responses generated by different LLMs, we can observe the level of agreement and infer the consistency in their outputs for a given prompt. Specifically, we hypothesized that using a majority voting approach from the ensemble of prompt responses would boost the reliability of language models, minimizing potential errors, variations, and biases associated with individual prompting approaches. To achieve this, in independent chats, we prompted the LLM with 3 distinct prompt formats per case, as presented in Table 1. Subsequently, we collected the responses of each model and applied majority voting to aggregate its predictions, as presented in Figure 1. In majority voting, each prompt produced a response from the language model, and the majority response was chosen as the final response. In a scenario where all prompt strategies resulted in different responses, we assumed that the model was unsure of that question and scored the final response as a failure case. We limited the number of prompts in the ensemble to 3 because studies by Wang et al [28] and Li et al [29] have shown that we obtain diminishing returns as we increase the overall number of prompts in an ensemble.

**Figure 1.** Our proposed method contains the following steps: (1) prompt a language model using a distinct set of prompts, (2) obtain diverse responses, and (3) choose the most consistent response as the final answer (majority voting).

**Ethical Considerations**

No ethics approval was pursued for this research, given that the data was publicly accessible and deidentified. This aligns with the guidelines outlined in the National Institutes of Health investigator manual for human subjects research [30].
Results

Performance Across Prompt Strategies

Figure 2 reveals the performance of LLMs across different prompts on the DC3 data set. Overall, approach 2 (multiple-choice prompt) yielded the highest score for all 30 cases, with GPT-4 and Bard achieving an accuracy score of 47% (14/30) and ChatGPT-3.5 obtaining a score of 43% (13/30). However, when considering case difficulty, the results varied.

On the frequently misdiagnosed cases category, GPT-4 and ChatGPT-3.5 performed better with open-ended prompts (approach 1), scoring 30% (3/10) and 20% (2/10), respectively. In contrast, Bard demonstrated superior performance with multiple-choice prompts for selection and ranking (approaches 2 and 3), achieving a score of 30% (3/10). ChatGPT-3.5 and Bard performed equally well on the rarely misdiagnosed cases category using approaches 2 and 3, achieving a perfect score of 100% (2/2). Furthermore, GPT-4 attained a score of 100% (2/2) but only with approach 2. For the moderately misdiagnosed cases category, all LLMs achieved their best performance with approach 2, scoring 67% (12/18), 56% (10/18), and 50% (9/18) for GPT-4, ChatGPT-3.5, and Bard, respectively. Table S1 in the Multimedia Appendix 1 presents the inconsistencies in the correct responses across the approaches for different cases. For example, Bard could only diagnose milk alkali syndrome using approach 1 but failed to use other prompt approaches. ChatGPT-3.5 correctly diagnosed primary adrenal insufficiency (Addison disease) with only approach 2, whereas GPT-4 was able to diagnose acute hepatitis E virus infection with only approach 1. These results indicate that no universal prompt approach is optimal for all LLMs when dealing with complex cases.

Results on the MIMIC-III data set in Figure 3 showed that the LLMs also performed best using approach 2 (multiple-choice prompt), with Bard and GPT-4 obtaining scores of 93% (14/15) each and ChatGPT-3.5 obtaining 73% (11/15). Using approach 3 (ranking prompt) resulted in a slight drop in performance for GPT-4 and Bard, with a 6% decrease, whereas the performance of ChatGPT-3.5 dropped by 26%. Approach 1 (open-ended prompt) proved challenging for the LLMs, with scores of 47% (7/15), 60% (9/15), and 27% (4/15) for Bard, GPT-4, and ChatGPT-3.5, respectively. Table S2 in the Multimedia Appendix 1 illustrates that approach 1 was only beneficial to GPT-4 in diagnosing amyloidosis, whereas it was consistently never the sole correct approach for Bard and ChatGPT-3.5. These results aligned with the findings from the DC3 data set and emphasized the varying performances of different models and prompt approaches across tasks.

Figure 2. Results of the diagnostic case challenge collection data set comparing prompt strategies. OpenAI GPT-4 outperformed all other models, achieving the highest score in all 30 cases using the majority voting approach. Furthermore, all large language models except MedAlpaca outperformed the human consensus (denoted by a black dashed line) across all cases, regardless of the difficulty, using at least 1 prompt approach. GPT-4: generative pretrained transformer-4.
Performance With Majority Voting

Previous experiments have demonstrated that there is no perfect prompting strategy because LLM users may not know beforehand which prompt will produce a correct response. We used the majority voting approach to estimate consistency, maximize the benefits of different prompt strategies, and enhance the reliability of the LLMs’ responses. Figure 2 illustrates the results for all DC3 cases. Majority voting improved the overall performance of GPT-4 from 47% to 50%, whereas the performance of ChatGPT-3.5 remained at 43% because majority voting did not decrease its performance compared with that of approach 2. In contrast, the performance of Bard decreased from 47% to 43% compared with that of approach 2. Summarizing the overall performance based on query difficulty, majority voting resulted in a perfect score of 100% for the rarely misdiagnosed cases category across all the LLMs. For the frequently misdiagnosed cases category in DC3, Bard achieved the highest score with majority voting and multiple-choice prompts, whereas GPT-4 performed best for the moderately misdiagnosed cases category with majority voting and approach 2. In addition, GPT-4 outperformed all other LLMs across all DC3 cases using the majority voting approach, regardless of the case difficulty. This score surpassed the performance of the individual prompt approaches in all cases.

Results on the MIMIC-III data set in Figure 3 showed that, the scores with majority voting were 87% (13/15) for GPT-4 and Bard each and 53% (8/15) for ChatGPT-3.5. These results indicate that the ensemble method did not substantially improve their performance compared with their best individual approach. It is worth noting that although the majority voting approach did not consistently outperform individual approaches in terms of the highest number of correct responses, it did provide a means to consolidate predictions and mitigate potential errors and biases from single approaches.

Comparison With Human Respondents

In the DC3 cases, although the human respondents had the advantage of accessing supporting patient information such as image scans and magnetic resonance imaging, the LLMs consistently outperformed the average human consensus. As shown in Figure 2, using the majority voting approach, all LLMs achieved a higher performance than the human consensus (denoted by a black dashed line), with a minimum margin of 5% across all 30 cases. Specifically, when considering query difficulty, the LLMs demonstrated even greater superiority. In the rarely misdiagnosed cases category, all LLMs surpassed the average human consensus by a substantial margin of 26%. For the moderately misdiagnosed cases category, GPT-4 and ChatGPT-3.5 maintained their advantage over human respondents, achieving a minimum margin of 11% with the majority voting approach. In contrast, only Bard outperformed the human average consensus on the frequently misdiagnosed cases category, with a margin of 14%.

We conducted a Spearman rank correlation test to analyze the pattern in the responses between each LLM and the human respondents. This involved correlating the average percentage of correct responses for each LLM across the prompt strategies with that of correct human responses. The results of the Spearman correlation test revealed that Bard had a relatively weak correlation coefficient of 0.30, whereas GPT-4 and ChatGPT-3.5 exhibited moderate positive correlations of 0.51 and 0.50, respectively. This suggested that the diagnostic performance patterns of GPT-4 and ChatGPT-3.5 aligned moderately with those of the human respondents. The observed correlation in answering patterns between human respondents and LLMs may stem from the inherent data bias present in the training data sets. The LLMs learn from vast amounts of data, and if the training data are biased toward certain diagnostic or decision-making patterns commonly expressed by human physicians, the model is likely to replicate those patterns. Although the correlation suggested that the LLMs have the
potential to be valuable tools in medical education, it is important to note their correlation with human physicians and that the performance of LLMs does not necessarily mean that they are as good as human physicians in diagnosing and treating diseases.

We could not directly compare the performance of human respondents on the MIMIC-III data sets because of the unavailability of data. Overall, the results indicated that the LLMs consistently outperformed the average human consensus in diagnosing medical cases, showcasing their potential as a tool to complement and enhance care quality and education for complex diagnostic cases.

Comparison With MedAlpaca

On the DC3 data sets, Bard, GPT-4, and ChatGPT-3.5 outperformed MedAlpaca across all cases using the majority voting approach by a minimum margin of 13%. MedAlpaca also displayed the worst performance in the open-ended prompts, irrespective of query difficulty. However, when multiple-choice options were provided, MedAlpaca outperformed the other LLMs in the frequently misdiagnosed cases category. Similar to the DC3 data set, MedAlpaca consistently demonstrated its best performance using the ranking prompt on the MIMIC-III data sets. However, its overall performance was significantly poorer than the other LLMs, with each LLM outperforming the model by at least 26% using the majority voting approach. In contrast to the general-purpose LLMs (eg, Bard, GPT-4, and ChatGPT-3.5), investigating the MedAlpaca model was finetuned using diverse medical tasks and assessed using multiple-choice medical examinations. This tailored training approach likely contributed to its notable performance, particularly excelling in DC3 cases (frequently misdiagnosed instances) and demonstrating optimal results in multiple-choice queries.

Qualitative Analysis

In our experiments, we manually observed the responses of each LLM to all our prompts and noted that each LLM consistently justified its diagnosis choice except for MedAlpaca. Specifically, each LLM offered a logical explanation for its chosen response regardless of the prompting strategy. For further investigation, we analyzed each LLM’s responses in 3 scenarios: (1) when presented with multiple-choice options containing the true diagnosis and they responded accurately, (2) when their response was incorrect, and (3) when given only incorrect multiple-choice options to pick from. In the first scenario, as presented in Multimedia Appendix 1, all 3 LLMs emphasized that the patient’s systemic symptoms, exposure risks, chest radiograph, computed tomography scan findings, and the suspected compromised ulcers were owing to relevant symptoms presented in the case, such as a history of respiratory illness and the presence of mesenteric lymph nodes and numerous tiny nodules throughout both lungs distributed in a miliary pattern. This pattern of offering insightful reasons for the likelihood of a diagnosis and explaining why other diagnostic options are less probable is valuable for educational purposes. In the second scenario, we observed that there was a notable disparity in the accuracy of human respondents. Only 6% (217/3624) of the human participants provided the correct response, with most votes (1232/3624, 34%) favoring ulcerative colitis, whereas 23% (833/3624) of the human responses opted for salmonellosis. Notably, Bard and GPT-4 displayed similar behavior by selecting salmonellosis, whereas ChatGPT-3.5 and MedAlpaca chose ulcerative colitis.

Another notable finding occurred in the responses of GPT-4 and ChatGPT-3.5. Regardless of the correctness of their chosen diagnoses, these models consistently recommended further tests to confirm their responses. This behavior suggested a general tendency toward advocating additional examinations to validate their diagnoses, potentially reflecting a cautious approach. In contrast, Bard adopted a different approach. Instead of recommending further tests, Bard highlighted that the provided query information supported the diagnosis without suggesting additional confirmatory measures. In the scenario where only incorrect options were given, Bard, ChatGPT-3.5, and MedAlpaca made choices and justified their responses. In contrast, GPT-4 explicitly mentioned that none of the provided options matched the case presentation. Furthermore, GPT-4 suggested a more probable diagnosis and recommended additional testing to explore its feasibility.

Discussion

Principal Findings

Previous studies [4,5] have presented the impressive success of LLMs in standardized medical examinations. We conducted experiments to assess the potential of LLMs as a CME system for rare and complex diagnoses, and our findings demonstrated that LLMs have the potential to be a valuable tool for rare disease education and differential diagnosis. Although LLMs demonstrated superior performance compared with the average human consensus in diagnosing complex diseases, it is essential to note that this does not imply their superiority over physicians. Numerous unknown factors, including the level of respondents’ expertise, may influence the outcome of web-based polls. Furthermore, we examined the knowledge capacity of LLMs through open-ended and multiple-choice prompts and found that LLMs, including MedAlpaca, performed better with multiple-choice prompts. This improvement can be attributed to the options provided, which narrowed the search space for potential diagnoses from thousands to a few likely possibilities. Consequently, we surmise that LLMs are not yet ready to be used as stand-alone tools, which aligns with the findings of previous studies [5,17,18]. Our observations revealed the consistent outperformance of general-purpose LLMs over MedAlpaca in various experiments. Their superior ability to provide valuable justifications for making diagnoses was particularly noteworthy, a strength not matched by MedAlpaca. This difference may stem from MedAlpaca’s exclusive finetuning and assessment for multiple-choice medical examinations, which slightly differ in format from the clinical cases in our experiments.

A notable finding in the response of LLMs to queries was their consistent provision of coherent and reasoned explanations, regardless of the query format. For instance, when diagnosing miliary tuberculosis, all 3 LLMs emphasized that the patient’s systemic symptoms, exposure risks, chest radiograph, computed tomography scan findings, and the suspected compromised...
immune state collectively support the diagnosis of *miliary tuberculosis*. Furthermore, Bard and GPT-4 ruled out other diagnoses presented in the multiple-choice prompt by highlighting their less typical presentations and lack of certain associated symptoms or risk factors. In addition, the conversational nature of LLMs allows users to ask follow-up questions for further context. These attributes hold great potential for educating users and offering them insights. However, we observed that LLMs provided logical explanations, even when their diagnoses were incorrect. ChatGPT-3.5 and GPT-4 may suggest additional testing to validate their selected diagnosis or use cautious terms like “potential diagnosis.” However, it remains unclear whether these recommendations stem from the models’ internal confidence or whether there are features intentionally designed by the developers for cautious use. The absence of explicit information regarding the level of uncertainty of LLMs for a specific case is concerning as it could potentially mislead clinicians. The ability to quantify uncertainty is crucial in medical decision-making, in which accurate diagnoses and treatment recommendations are paramount. Clinicians heavily rely on confidence levels and probability assessments to make informed judgments [29]. Without an indication of uncertainty, there is a risk that clinicians may trust the logical explanations provided by the LLMs even when they are incorrect, leading to misdiagnoses or inappropriate treatment plans.

Considering the delicate role of clinical decision support, it is essential to address validity and reliability as crucial aspects of uncertainty. Moreover, a reliable system is of paramount importance for medical education. However, the stochastic nature of LLMs introduces doubts among clinicians regarding their reliability. Although a specific metric to quantitatively assess the reliability of the LLMs used in this study is currently lacking, we acknowledge the significance of consistency in achieving reliability. To address this, we used different prompting strategies and implemented a majority voting approach to select the most consistent response from each LLM. After examining the individual prompt strategies, we anticipated consistent responses across strategies for a specific case. However, our findings revealed that the responses of LLMs were sensitive to concrete prompt formats, particularly in complex diagnoses. For instance, ChatGPT-3.5 and GPT-4 performed better with the open-ended prompt (approach 1) in the frequently misdiagnosed cases category of DC3 cases but struggled with similar cases using multiple-choice and ranking prompts (approaches 2 and 3). In contrast, Bard performed better with multiple-choice prompts. These results highlighted that there is no one-size-fits-all prompting approach nor does a single strategy apply universally to all LLMs. Although the majority voting strategy did not yield optimal results for all models across data sets, it served as a means to consolidate responses from multiple prompts and provided a starting point for incorporating reliability.

Several studies [10-12,14,15] have emphasized the significance of enhancing the education of clinicians at all levels to provide better support for rare and complex diagnoses. In this pursuit, the studies by Lee et al [8] and Decherchi et al [31] have highlighted the potential advantages of artificial intelligence (AI) systems, whereas the studies by Abdullahi et al [25] and Sutton et al [32] have reported a lack of acceptance of AI tools among clinicians. For instance, younger medical students and residents appeared more receptive to integrating technology [33]. One notable reason for this lack of acceptance is that conventional AI systems typically require training before clinicians can effectively use them, which can be burdensome and time consuming [32]. In contrast, conversational LLMs, such as ChatGPT-3.5, Bard, and GPT-4, offer a distinct advantage with their simple interface and dialogue-based nature. These conversational LLMs eliminate the need for extensive training, increasing their potential for high acceptance across all levels of medical practice. Although the exciting ease of use, conversational nature, impressive display of knowledge, and logical explanations of LLMs have the potential for user education and insights, their current limitations in reliability and expressing uncertainty must be addressed to ensure their effective and responsible use in critical domains, such as health care.

**Limitations**

First, the limitations of the knowledge of ChatGPT-3.5 and GPT-4 to the latest trends and updates in health care (or medical) data till 2021 pose the risk of potentially incomplete information and hamper the effectiveness of the models as a CME tool, especially when addressing emerging diseases. In contrast, although continuous updates to Bard are advantageous for keeping the model up-to-date, this attribute may impact the reproducibility of our study. Second, it is notable that our experiments had a limited scope owing to a small sample size consisting of only 30 diseases from the DC3 data set and 15 cases from the MIMIC-III data set. In addition, although we took precautions to preprocess the MIMIC-III notes to prevent leakage of the final diagnosis, the discharge summaries may still contain nuanced information that could make the diagnosis obvious. Furthermore, the closed nature of the LLMs used in this study restricted our technique for measuring reliability to a majority voting approach, which consolidated responses from diverse prompts. Although majority voting can help to mitigate the variability of LLM output, it is notable that LLMs may still generate different responses for the same prompt. This variability should be considered when interpreting the results of this study. However, when these LLMs are released with an enhanced iteration that allows for finetuning and calibration, future work should incorporate more effective mechanisms to estimate and communicate uncertainty. An example of such an approach could involve assigning a confidence score to the probability score of their responses. This methodology could allow clinicians to make informed decisions regarding whether to accept or reject responses that fall within a desired threshold.

**Conclusions**

In this study, we conducted experiments to assess the potential of LLMs, including ChatGPT-3.5, GPT-4, and Bard, as a CME system for rare and complex diagnoses. First, we evaluated their diagnostic capability specifically for rare and complex cases. Subsequently, we explored the impact of prompt formatting on their performance. Our results revealed that these LLMs possessed potential diagnostic capacities for rare and complex...
medical cases, surpassing the average crowd consensus on the DC3 cases. For selected rare cases from the MIMIC-III data set, Bard and GPT-4 achieved a diagnostic accuracy of 93%, whereas ChatGPT-3.5 achieved an accuracy of 73%. Our findings highlighted that users might discover an approach that yields favorable results for various queries by exploring different prompt formats. In contrast, using majority voting of responses from multiple prompt strategies offers the benefit of a robust and reliable model, instilling confidence in the generated responses. However, determining the best prompt strategy versus relying on the majority voting approach involves a tradeoff between exploration and exploitation. Although prompt engineering research is continuing, we hope that future studies will yield better solutions to enhance the reliability and consistency of the responses of LLMs. Overall, our study’s results and conclusions provide a benchmark for the performance of LLMs and shed light on their strengths and limitations in generating responses, expressing uncertainty, and providing diagnostic recommendations. The insights gained from this study can serve as a foundation for further exploration and research on using LLMs as medical education tools to enhance their performance and capabilities as conversational language models.

Acknowledgments
We acknowledge support from the Open Access Publication Fund of the University of Tübingen.

Data Availability
The URLs for the diagnostic case challenge collection data set can be obtained via A Diagnostic Case Challenge Collection [34]. The Medical Information Mart for Intensive Care data sets can be accessed via the database, Medical Information Mart for Intensive Care-III Clinical Database v1.4 [35], after obtaining permission from Physionet.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Comprehensive tables detailing the performance of each model across data sets, with included examples of prompts and responses for each model.

References
1. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-03-23]


Abbreviations

AI: artificial intelligence
CME: continuing medical education
DC3: diagnostic case challenge collection
LLM: large language model
MIMIC-III: Medical Information Mart for Intensive Care-III
USMLE: United States Medical Licensing Examination
Using ChatGPT-Like Solutions to Bridge the Communication Gap Between Patients With Rheumatoid Arthritis and Health Care Professionals

Chih-Wei Chen\(^1,2,3,4^{*}\), MPhil; Paul Walter\(^1,5,6\), MSc; James Cheng-Chung Wei\(^3,7^{*}\), MD, PhD

\(^1\)National Applied Research Laboratories, Taipei, Taiwan
\(^2\)National Council for Sustainable Development, Taipei, Taiwan
\(^3\)Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan
\(^4\)Faculty of Engineering Sciences, University College London (UCL), London, United Kingdom
\(^5\)Faculty of Pharmacy, Paris-Saclay University, Orsay, France
\(^6\)Mines Saint-Etienne, Saint-Etienne, France
\(^7\)Department of Allergy, Immunology & Rheumatology, Chung Shan Medical University Hospital, Taichung, Taiwan

* these authors contributed equally

Corresponding Author:
Chih-Wei Chen, MPhil
National Applied Research Laboratories
3F, No 106, Sector 2
Heping East Road
Taipei, 106214
Taiwan
Phone: 886 975303092
Email: chihwei.chen@udm.global

Abstract

The communication gap between patients and health care professionals has led to increased disputes and resource waste in the medical domain. The development of artificial intelligence and other technologies brings new possibilities to solve this problem. This viewpoint paper proposes a new relationship between patients and health care professionals—“shared decision-making”—allowing both sides to obtain a deeper understanding of the disease and reach a consensus during diagnosis and treatment. Then, this paper discusses the important impact of ChatGPT-like solutions in treating rheumatoid arthritis using methotrexate from clinical and patient perspectives. For clinical professionals, ChatGPT-like solutions could provide support in disease diagnosis, treatment, and clinical trials, but attention should be paid to privacy, confidentiality, and regulatory norms. For patients, ChatGPT-like solutions allow easy access to massive amounts of information; however, the information should be carefully managed to ensure safe and effective care. To ensure the effective application of ChatGPT-like solutions in improving the relationship between patients and health care professionals, it is essential to establish a comprehensive database and provide legal, ethical, and other support. Above all, ChatGPT-like solutions could benefit patients and health care professionals if they ensure evidence-based solutions and data protection and collaborate with regulatory authorities and regulatory evolution.

(JMIR Med Educ 2024;10:e48989) doi:10.2196/48989

KEYWORDS
rheumatoid arthritis; ChatGPT; artificial intelligence; communication gap; privacy; data management

Introduction

In recent years, the communication gap has led to intense relationships between patients and health care professionals. The use of ChatGPT-like solutions in health care has enormous potential to improve the patient-provider relationship, such as patient clinic letter writing [1], medical note-taking and consultation [2], and rheumatoid arthritis treatment. Although ChatGPT (OpenAI) [3] is not the only solution available, the technology has generated a lot of traction due to its advanced features, such as the ability to enhance rule-based chatbots. However, it is important to note that ChatGPT-like solutions should not be viewed as a stand-alone solution but as an integrated interface in a larger ecosystem that allows access to
multiple data sources. In terms of the patient-provider
relationship, the use of ChatGPT can enable more fluid and
effective communication between the 2 parties, which can
improve the quality of care. In particular, the use of
ChatGPT-like solutions in the context of methotrexate treatment
for rheumatoid arthritis could have a significant impact. This
viewpoint paper proposes a new relationship between patients
and providers—“shared decision-making”; explains the potential
of ChatGPT-like solutions in improving the patient–health care
professional relationship from the clinical and patient
perspectives; and suggests the importance of establishing a
comprehensive database to promote the implementation of
“shared decision-making” between patients and health care
professionals.

**Toward Shared Decision-Making**

In conventional medical settings, the relationship between
patients and health care professionals was not equal, mainly
because of the huge information gap between them, since
patients lacked medical knowledge and decision-making
capacity. In recent years, the rapid development of ChatGPT
possesses enormous potential to bridge the information gap and
improve the relationship between patients and health care
professionals. For instance, ChatGPT could help provide the
risk-benefit analysis of different treatment options, assisting
health care professionals and patients to understand the
advantages and disadvantages of each option and then make
informed decisions together. It could also assist patients in
understanding the complex medical jargon and technical details
and provide information about the disease, treatment options,
potential risks, and expected outcomes, allowing patients to
participate in making informed decisions with health care
professionals together. ChatGPT-like solutions allow bilateral
communications between patients and health care professionals
toward shared decision-making.

**Clinical Perspective**

From the clinical standpoint, early diagnosis of rheumatoid
arthritis is crucial [4] for health care professionals and should
be based on clinical examinations and biological results such
as serological tests [5]. However, the differential diagnosis of
complex diseases such as rheumatoid arthritis–associated
interstitial lung disease [6] remains a major concern [7], as it is
responsible for a significant increase in mortality [8].
ChatGPT-like solutions could bring complementary support to
diagnose the disease and predict its evolution. Thus, to the query
“What could be the reason for cough and dyspnea in a patient
with rheumatoid arthritis?” ChatGPT suggests interstitial lung
disease in the first place. By integrating external data on risk
factors (age, sex, and smoking), biological results (pulmonary
function testing, autoantibodies, and biopsy), and imaging
(high-resolution computed tomography and ultrasound) [6],
ChatGPT-like solutions can assist in suggesting additional tests
and confirming the diagnosis.

The initiation of treatment for rheumatoid arthritis should be in
accordance with the latest official recommendations, such as
those from the European League Against Rheumatism [9] and
the American College of Rheumatology (ACR) [4]. An advanced
tool such as ChatGPT provides clinicians with exhaustive
information on the latest guidelines for the management of
rheumatoid arthritis. For instance, if a clinician asks “What are
the current guidelines for treating rheumatoid arthritis according
to the ACR?” ChatGPT can retrieve the key points of
rheumatoid arthritis management in accordance with the official
ACR guidelines [4]. However, in the specific case of a request
regarding recommendations for treating rheumatoid arthritis–associated interstitial lung disease, ChatGPT erroneously refers to nonexistent ACR guidance [10]. Currently, the tool has limitations, such as data exclusion after 2021 and response size limits.

The determination of a patient’s drug dose by the clinician is based on a comprehensive evaluation of the results of the
biological tests and clinical examination. However, dose
adjustment may not always be performed according to a
standardized procedure and evidence-based solution, although
this is crucial to ensure the effectiveness and tolerability of the
treatment for the patient. Methotrexate is the most common
treatment for rheumatoid arthritis, and an initial dose of 7.5-15
mg once a week is recommended, followed by a gradual increase
in dose. However, poor patient adherence and nonpersistence
to methotrexate therapy have been reported [11] mainly due to
low dose tolerance. Optimization of methotrexate dose is therefore essential for treating rheumatoid arthritis [12].
The use of methotrexate monotherapy has shown similar efficacy
to the combined use of methotrexate monotherapy with biologic
disease–modifying antirheumatic drugs [13]. Process automation
and integration of complementary data, based on solutions such as
ChatGPT, could improve outcome prediction, contribute to
drug dose optimization, and thus reduce costs to the health care
system.

Access to information on ongoing clinical trials and their results
would enable clinicians to propose treatments for people with
rare conditions in rheumatoid arthritis. Compiling data on
clinical trials and patient characteristics would allow clinicians
to propose alternatives, for example, for patients who have failed
current therapies. Identifying subpopulations would facilitate
patient recruitment and bring more effective and safer drugs to
market. However, one challenge is to deidentify data to comply
with the US Health Insurance Portability and Accountability
Act (HIPAA) [14]. As such, it is important for clinicians to
prioritize patient privacy and confidentiality when accessing
and using such data. In addition, it is necessary for further
interdisciplinary research to improve the accuracy and
persuasiveness of artificial intelligence (AI) chatbots to influence
patients’ behaviors [15]. Moreover, the application of AI and
machine learning in health care should still be regulated by
establishing norms to reduce bias and reflect the real problems
[16].

**Patient Perspective**

From the patient’s perspective, it allows easy access to a large
volume of information with a certain degree of scientific
evidence, which improves the patient’s knowledge of
rheumatoid arthritis and their health literacy. ChatGPT-like
solutions thus contribute to dealing with the proliferation of unreliable sources of emerging information and widespread disinformation [17]. It is also a tool that could not only enable empowerment by acting interactively throughout the care pathway but also promote patient adherence to treatment. However, some concerns persist regarding the lack of supervision of this type of solution and the liability involved [18]. For example, in the case of methotrexate side effects, to the query “I have gastrointestinal problems and fatigue, is this related to my methotrexate intake?” ChatGPT suggests that the doctor can adjust the dosage. It does not provide suggestions to state that concomitant folic acid would reduce toxicity. It also raises questions about the risk of patients adjusting their own dosage. ChatGPT-like solutions can strengthen expert patients’ collaboration, allowing the cocreation of care pathways; however, it can also be a source of conflict by pitting the tool’s and the caregiver’s advice against each other. Therefore, it is crucial to better supervise this tool from the beginning of its development, in order to clearly distinguish between its general public and medical use and to define the responsibilities of each. The use of ChatGPT-like solutions can improve communication and access to information for patients with rheumatoid arthritis but must be carefully managed to ensure safe and effective care.

A ChatGPT-like solution allows the patient to have continuous access to information in an interactive way that promotes understanding outside the clinical setting. This solution can play an important role in therapeutic education by providing information on the self-management of rheumatoid arthritis, on a drug such as methotrexate, or on the administration methods (oral and subcutaneous). Therefore, the query of “What precautions should be taken when taking methotrexate?” could instantly provide basic and exhaustive information (taking it with food, avoiding alcohol, staying hydrated, using contraceptives, etc) and could contribute to therapeutic education [19]. In addition, a ChatGPT-like solution could be used to communicate medical information on potential benefits and assist in administration [20], for example, when modifying the route of administration of methotrexate. This would have an impact on facilitating the acceptance of subcutaneous methotrexate, allowing better bioavailability and clinical efficacy. It would also reduce the time required to initiate treatment and avoid the use of biologics, thus having a significant impact on health care costs [21].

Further integration and analysis of patient requests would also accelerate the transition to more personalized medicine. ChatGPT-like solutions could identify patient profiles and adapt communication strategies to overcome resistance and nudge behavior. These solutions will have to be adapted to each country in terms of public health systems and beliefs.

Establishment of a Comprehensive Database

The database is one of the critical elements of digital infrastructure for digital technology applications [22], especially AI-based solutions that require huge amounts of data to achieve more accurate results. However, using AI-based technology can be limited by the nontransparent learning process, difficulties in explanation and validation, and the influence of improper data [23]. Hence, the establishment of a comprehensive database, which is sourced from real-world data and updated on time and precisely, could contribute to overcoming limitations caused by insufficient data and support evidence-based clinical applications.

In recent decades, the Taiwan government launched the National Health Insurance (NHI) system that collected health-related data of health care providers, citizens, and legal residents. Since its establishment, the NHI database has been continuously improved by using the latest technologies to accommodate the increasing needs. During the COVID-19 pandemic, the NHI database successfully supported the Taiwan government in tracking patients, distributing face masks, and containing the infections [24,25].

On the other hand, using mobile health tools also contributes to the establishment of a comprehensive database. In recent years, tools such as the Apple Watch have been widely used to collect data about health conditions and identify possible illnesses of people. Mobile health tools allow the collection of active and passive data, which could better inform the health condition of the people [26].

Above all, the establishment of a comprehensive database is fundamental to applying AI-based solutions in the digital governance of health care. Moreover, applying AI-based solutions and other digital technologies should also be accompanied with comprehensive planning and flexible strategies to achieve effective digital governance in health care [22].

Conclusions

In conclusion, ChatGPT-like solutions have the potential to improve the patient-provider relationship through “shared decision-making.” ChatGPT solutions should optimize the patient’s care pathway while improving the patient’s experience of using methotrexate in rheumatoid arthritis. However, there is a need to ensure evidence-based solutions and quantify these benefits. In the future, we may question the compatibility of the business model of mass-market solutions with health care system purposes, particularly concerning data protection. Using federated learning might be a way for developers to overcome this limitation. The implementation in a specific health care context should increase in the coming years with the development of solutions in specific domains such as Bio-Generative Pre-Trained Transformer. A deployment in clinical settings will require collaboration with regulatory authorities and potentially an evolution of the software as a medical device regulatory framework [27].

The need to include individuals in the design of these solutions is also crucial to consider from an efficiency point of view to avoid certain biases and from an ethical point of view. This solution also facilitates access to health care information for the entire world population in pursuit of the sustainable development goals set by the United Nations.
None declared.

References


Abbreviations

ACR: American College of Rheumatology
AI: artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
NHI: National Health Insurance

©Chih-Wei Chen, Paul Walter, James Cheng-Chung Wei. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Exploring the Feasibility of Using ChatGPT to Create Just-in-Time Adaptive Physical Activity mHealth Intervention Content: Case Study

Abstract

Background: Achieving physical activity (PA) guidelines’ recommendation of 150 minutes of moderate-to-vigorous PA per week has been shown to reduce the risk of many chronic conditions. Despite the overwhelming evidence in this field, PA levels remain low globally. By creating engaging mobile health (mHealth) interventions through strategies such as just-in-time adaptive interventions (JITAs) that are tailored to an individual’s dynamic state, there is potential to increase PA levels. However, generating personalized content can take a long time due to various versions of content required for the personalization algorithms. ChatGPT presents an incredible opportunity to rapidly produce tailored content; however, there is a lack of studies exploring its feasibility.

Objective: This study aimed to (1) explore the feasibility of using ChatGPT to create content for a PA JITAI mobile app and (2) describe lessons learned and future recommendations for using ChatGPT in the development of mHealth JITAI content.

Methods: During phase 1, we used Pathverse, a no-code app builder, and ChatGPT to develop a JITAI app to help parents support their child’s PA levels. The intervention was developed based on the Multi-Process Action Control (M-PAC) framework, and the necessary behavior change techniques targeting the M-PAC constructs were implemented in the app design to help parents support their child’s PA. The acceptability of using ChatGPT for this purpose was discussed to determine its feasibility. In phase 2, we summarized the lessons we learned during the JITAI content development process using ChatGPT and generated recommendations to inform future similar use cases.

Results: In phase 1, by using specific prompts, we efficiently generated content for 13 lessons relating to increasing parental support for their child’s PA following the M-PAC framework. It was determined that using ChatGPT for this case study to develop PA content for a JITAI was acceptable. In phase 2, we summarized our recommendations into the following six steps when using ChatGPT to create content for mHealth behavior interventions: (1) determine target behavior, (2) ground the intervention in behavior change theory, (3) design the intervention structure, (4) input intervention structure and behavior change constructs into ChatGPT, (5) revise the ChatGPT response, and (6) customize the response to be used in the intervention.

Conclusions: ChatGPT offers a remarkable opportunity for rapid content creation in the context of an mHealth JITAI. Although our case study demonstrated that ChatGPT was acceptable, it is essential to approach its use, along with other language models, with caution. Before delivering content to population groups, expert review is crucial to ensure accuracy and relevancy. Future research and application of these guidelines are imperative as we deepen our understanding of ChatGPT and its interactions with human input.

(JMIR Med Educ 2024;10:e51426) doi:10.2196/51426
KEYWORDS
ChatGPT; digital health; mobile health; mHealth; physical activity; application; mobile app; mobile apps; content creation; behavior change; app design

Introduction

Physical inactivity is a key modifiable risk factor for many chronic conditions, including cardiovascular disease, type 2 diabetes, and cancers, throughout the lifespan [1]. Despite this evidence, adults and adolescents alike are not consistently meeting the recommended guidelines to prevent developing these chronic conditions [2]. Previous studies have shown that 150 minutes of moderate-to-vigorous physical activity (MVPA) can reduce the risk of all-cause mortality by at least 30%, along with reducing the risk for chronic conditions such as cardiovascular disease (30%), colon cancer (20%), and breast cancer (14%) [3]. Although many chronic diseases affect adults, healthy lifestyle habits need to be developed early from childhood. Children aged 8 to 12 years are more flexible than adults in their ability to change behaviors because they are just beginning to develop self-regulation skills, habits, and identities for healthy living [4,5]. Thus, many countries such as Canada [6], the United States [7], and the United Kingdom [8] have set guidelines recommending 60 minutes of MVPA per day for children 17 years and younger [2]. However, despite these recommendations, physical inactivity is prevalent among children, with less than one-quarter of children meeting the guidelines in countries such as Canada [9] and the United States [9]. Consequently, promoting regular PA to prevent chronic diseases and maintain lifelong health has been a key priority for governments worldwide.

Recent studies suggest that family-based PA programs can be an effective strategy to improve PA levels in children [10,11]. These programs focus on providing guidance for parents to support their child’s PA (eg, encouragement, providing opportunity, and logistic support) [12]. With advancements in mobile health (mHealth) technologies and improved access to smartphones, emerging evidence indicates that PA interventions delivered through mHealth technology can be effective while improving scalability and personalization. However, the effectiveness and engagement of interventions vary depending on the intervention design and the degree of tailoring [13,14]. Studies have demonstrated that tailored mHealth interventions are more effective in improving behavior and health outcomes compared with nontailored interventions [15]. A recent advancement in tailored mHealth interventions is the development of just-in-time adaptive interventions (JITAIs), which use mHealth technology to assess the dynamically changing needs of individuals and deliver tailored support in real time [14,16]. Thus far, JITAIs have shown great promise in promoting PA among adults [17], university students [18], and chronic disease populations [19]. Further, innovative mHealth “no-code” development platforms, such as Pathverse, have made the development and implementation of JITAIs much easier and cost-effective [20,21]. However, the development of content for JITAIs can be extremely labor-intensive due to the need to create various versions of health-related content for different tailored algorithms. Although the documentation of content creation timelines for PA JITAIs is in its infancy, a typical timeline for PA content creation from the formative phase to pilot testing reportedly ranges from 12 [22] to 15 months [23,24].

Specific to JITAIs, the typical process of creating evidence-based and engaging content for these mHealth interventions typically involves the following steps [21,25]: (1) defining the behavior change theories and behavior change techniques (BCTs) required for the intervention [26]; (2) gathering evidence from various sources, such as previous literature, public health sources, gray literature, and blogs, and then adapting it to suit the needs of the intervention and deliver it through the chosen medium; and (3) writing content that is engaging and matches the literacy level of the target population for the app. These steps can often be time-consuming, with the need for researchers to follow these steps iteratively and repetitively for the duration of the design of the intervention. Further, despite the consideration of these steps, several challenges still arise in the development of JITAIs content. Existing studies have identified limitations, such as the need for more extensive content within interventions, struggles in creating novel and meaningful messages, and challenges in tailoring messages to diverse user preferences [27,28]. These studies have also recognized the resource constraints in developing content to meet these needs and the complex, multidimensional nature of creating tailored and engaging content for their sample. Therefore, an artificial intelligence (AI) tool such as ChatGPT (OpenAI) [29] can be extremely useful in making the process of generating JITAIs content for mHealth interventions faster and more cost-effective. ChatGPT offers a solution to the need for more content within interventions by leveraging its vast training data and the ability to generate a diverse set of messages efficiently. Further, its generative capabilities and the ability for users to continually prompt new rules address the challenge of creating novel content, reducing the risk of messages being perceived as overly simplistic.

ChatGPT was first launched by OpenAI in November 2022 and is an open AI language model that generates human-like responses to text-based prompts [30]. It can understand and generate responses in various languages, as well as debug code, write stories in different genres and lengths, summarize information from complex texts, offer explanations on various topics, and even reject answering inappropriate prompts [31]. Unlike other generative large language models (LLMs), ChatGPT stands out as the inaugural member of a series of highly scaled LLMs that attain state-of-the-art performance with minimal need for fine-tuning [32]. Further, ChatGPT is highly sophisticated in that it is able to provide continuous dialog by remembering what the user has said earlier in the conversation thread [33].

Although ChatGPT hosts an impressive suite of features and capabilities, there are also several ethical and privacy concerns
to keep in mind while using this service. First, it is important to note that ChatGPT “learns” its information from human input. This is subject to error and is limited based on what others have input into its system. Further, when generating health information content, in particular, this LLM has been extensively trained with data up to 2021, thus limiting some of the relevance and accuracy of current practices [34]. Second, ChatGPT stores its data in the United States, which, depending on the type of information being input into the United States, may be subjected to privacy concerns based on US freedom and privacy laws. To build on this consideration of data storage, it is crucial not to input any personal health information or other sensitive data into ChatGPT, as this LLM continues to learn from text prompts.

Since its inception, ChatGPT has been widely cited in various bodies of behavioral science literature as a virtual assistant, chatbot, and language translation tool [35]. To generate output from the program, a concept called prompt engineering is one method that explains how ChatGPT generates output [36]. In LLMs, a prompt is defined as an instruction to the model that customizes, enhances, or refines the output [37]. However, there is currently a lack of studies examining the feasibility of using ChatGPT to help develop intervention content for JITAIs aimed to promote PA when given a behavior change theory and a behavior target outcome.

Thus, the primary objective of this paper was to present an autoethnographic case study that explored the feasibility, including the acceptability and ease of use, of using ChatGPT to create content for a family-based PA JITAI mobile app. The secondary objective was to describe lessons learned and future recommendations for using ChatGPT in developing mHealth intervention content.

**Methods**

**Study Design**

This case study consisted of 2 phases, which took place from March 1, 2023, to April 30, 2023. In phase 1 (0-2 months), we used ChatGPT-3 to develop a 10-week family-based PA JITAI. In phase 2 (3-4 months), we described lessons learned based on our experience of using ChatGPT in phase 1 and provided future recommendations for using ChatGPT in the development of mHealth interventions.

**Ethical Considerations**

This paper outlines the procedural aspects of using ChatGPT for content generation for a subsequent study. Given that it operates independently without involvement of human participants or sensitive data, formal ethics approval from our institution was deemed unnecessary.

**Phase 1**

We explored the feasibility of using ChatGPT to create content for the PA JITAI mobile app. To determine the feasibility of using ChatGPT to rapidly create JITAI content, we used an autoethnographic case study approach [38]. This method enabled the researchers (AW and SL) to reflect on their experience of using ChatGPT. While using ChatGPT, the researchers created field notes and had a meeting to discuss their independent experiences with using ChatGPT-3. Specifically, we reflected on the acceptability and ease of use as key areas of focus for feasibility [39]. Results of the meeting were themed into acceptability and ease of use of using ChatGPT. Assessing acceptability metrics involved reflecting on the satisfaction of the response generated by ChatGPT. The ease-of-use assessment involved reflecting on ChatGPT usability [39]. In this phase, we used 2 tools, Pathverse and ChatGPT. Pathverse is a no-code app builder platform that supports mHealth research [20,40]. It consists of a web portal for researchers to create engaging mobile app interventions with “drag and drop” features instead of coding. The content is then instantly displayed on the Pathverse mobile app. We used ChatGPT-3 to generate the content needed to be added to Pathverse. To gather feasibility data, we generated intervention content to support parents to help their child (8-12 years of age) to be physically active.

The content generated for this app was developed based on the Multi-Process Action Control (M-PAC) framework [41,42]. The M-PAC framework addresses the intention-behavior gap through the understanding that ongoing reflective processes (ie, affective attitude and perceived opportunity) and regulation processes (ie, behavioral and cognitive tactics to maintain intention focus) are necessary for the intention to become an action [41]. Specific to a JITAI, the M-PAC framework was selected as the framework for this intervention to dynamically and contextually address users’ failed intentions to be physically active. Thus, the just-in-time intervention options can be tailored to the specific circumstances of the individual, aligning with either the reflective, regulatory, or reflexive process [41,42]. The M-PAC framework was additionally chosen as we have seen success with this framework and its associated BCTs (ie, action planning, repetition, and habit formation) in previous family-based PA programs [43]. To address these circumstances, our research team created decision tree algorithms to tailor the family lessons and challenges recommended throughout the weeks. The algorithms were designed using the M-PAC framework and take into consideration (1) child MVPA minutes, (2) parent support behavior, and (3) parent self-efficacy and motivation for supporting their child’s PA (Figure 1). Based on the decision tree, weekly tailored lessons needed to be created to target each M-PAC construct. Topics included parental support, affective attitudes toward supporting their children’s PA, capability, opportunity, self-monitoring of PA, and restructuring the environment for PA. These topics stemmed from previous research for family-based PA interventions using the M-PAC framework [43]. With these considerations, a variety of prompts were created based on these topics.
Figure 1. Names of modules in the decision tree algorithm for personalized lessons.

There are various components to consider when generating a prompt for ChatGPT. Specific to academic uses of ChatGPT, the elements to be included in a prompt include an instruction (ie, an overview of the output you would like to receive), context (ie, other background information to help tailor the output), input data (ie, additional specifications for the output that may include its strengths or limitations), and output indicator (ie, how you would like the output to be presented, including word count and paragraph format) [44]. When creating a prompt for this case study, we included the target behavior and the M-PAC framework, with each output to be delivered in bullet point form. Once the content was created, we then used the Pathverse mHealth no-code app design tool to develop the JITAI app [20,38-40].

Phase 2

We summarized lessons learned and future recommendations for using ChatGPT in the development of mHealth interventions. Our team identified common themes and patterns emerging from the process of creating the JITAI content using ChatGPT. We then compared our data with previous literature to develop recommendations for future use. This involved a literature search to identify relevant studies and lessons learned from using ChatGPT in mHealth interventions. The primary aim of the literature search was to gather a wide range of insights into the acceptability, including the application of ChatGPT and its effectiveness in this context and challenges associated with integrating ChatGPT into mHealth interventions to refine our recommendations.

Results

Phase 1: Exploring the Feasibility of Using ChatGPT to Create JITAI Content

The results of phase 1 are first reported on how the researchers (AW and SL) used ChatGPT to generate content, followed by an analysis of the feasibility of the use of ChatGPT in this context. Overall, we created 13 lessons with the help of ChatGPT in phase 1. Figure 2 displays an example of how this content was displayed in the mobile app. We provided specific prompts about the length of the content generated, the target constructs of the M-PAC framework, the tone of the lesson, and the literacy levels needed. We used multiple question prompts to optimize text output. Table 1 provides examples of prompts used for different lessons. We started with broad prompts (eg, explain the various constructs in the M-PAC framework) and then used specific prompts based on the output (eg, provide specific fun examples to help parents improve opportunities to support child PA; Table 1). After the prompts were input into ChatGPT, the output was copied into a separate document for review by the researchers (AW and SL). If more or alternate content was needed, prompts such as “provide additional information about [this topic]” were used. To ensure that the output given by ChatGPT was relevant and accurate, we referred to previous literature and previous content examples following the M-PAC framework [21,45,46]. Once the content was deemed acceptable and accurate by the researchers, it was uploaded to the Pathverse platform. This step additionally involved creating graphics to include along with the text responses and formatting the content into different app “pages” with fewer than 400 characters per page of the mobile app.
We evaluated the acceptability of ChatGPT for creating mHealth content by reflecting on content accuracy, relevance, and tone. Both researchers found that ChatGPT demonstrated an acceptable level of accuracy and relevance and provided relevant responses to the prompts. However, on some occasions, ChatGPT provided false academic references. This is a serious issue that needs to be addressed to prevent misinformation. Thus, both authors reflected the need to place a filtering mechanism to ensure that the content generated was appropriate. Furthermore, some of the answers lacked specificity (eg, provide examples of PA programs in my area). This may be due to the fact that ChatGPT-3 was trained using data up to September 2021. Finally, we found the tone of ChatGPT responses to be acceptable for research purposes. The overall tone matched the prompt given (eg, write in a fun and positive voice). Overall, ChatGPT did not generate any inappropriate content. There is an evident need to provide clear prompts in order for ChatGPT to provide optimal responses. Additionally, multiple questions are often needed to optimize ChatGPT responses. The researchers additionally agreed that providing a role to ChatGPT, for example, telling the LLM that it is a health researcher delivering a family-based PA intervention, may have further refined the tone and quality of the response given.

When reflecting on the feasibility of implementing ChatGPT for this case study, we (AW and SL) found ChatGPT to be easy to use. Both researchers (AW and SL), with varying levels of technical expertise, found the user interface to be intuitive. The ease of use also allowed us to test various prompts to help optimize the ChatGPT responses. Overall, we found that minimal training or prior experience is needed to use this tool, and it has the potential to make it widely accessible for researchers.

Figure 2. Screenshots of physical activity content generated by ChatGPT in the mobile app Pathverse.
Table 1. Prompts and responses from a conversation with ChatGPT in generating content for a 10-week family-based physical activity mobile app.

<table>
<thead>
<tr>
<th>Prompt entered</th>
<th>ChatGPT response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write a fun to read article about the role of parental capabilities to support child physical activity based on the M-PAC framework</td>
<td>• Are you struggling to find ways to support your child’s physical activity? Do you want to improve your parental capabilities to make physical activity a fun and regular part of your child’s life? Look no further than the M-PAC framework, which provides a roadmap for parents to support their child’s physical activity. The M-PAC framework stands for Multi-Process Action Control, and it emphasizes the importance of three key components for parents to successfully support their child’s physical activity: instrumental support, regulatory support, and motivational support. Instrumental support refers to the tangible resources and skills that parents provide to support their child’s physical activity. This could include things like providing transportation to and from activities, purchasing sports equipment, or enrolling your child in a physical activity program. Regulatory support focuses on creating an environment that supports physical activity. This could include setting rules and boundaries around screen time, creating a routine for physical activity, or providing healthy food options to fuel physical activity. Motivational support is all about fostering a positive attitude towards physical activity. This could include providing praise and encouragement, setting goals, and finding ways to make physical activity fun and enjoyable.</td>
</tr>
<tr>
<td>Provide specific fun examples to help parents improve their capabilities for supporting child physical activity in bullet points</td>
<td>• Certainly! Here are some specific fun examples to help parents improve their capabilities for supporting child physical activity: 1. Instrumental Support: • Have a weekly family game night that involves physical activity such as charades, Pictionary or dance party. • Organize a scavenger hunt with physical activity challenges for your child and their friends. • Plan a family hike or outdoor adventure at a nearby park or nature reserve. 2. Regulatory Support: • Start a vegetable garden together and teach your child the importance of physical activity and nutrition for a healthy lifestyle. • Use a fitness tracker or pedometer with your child to encourage them to walk more and set goals for daily steps. • Plan a family bike ride or roller skating adventure on a weekend afternoon. 3. Motivational Support: • Make a fun reward system for reaching physical activity goals, such as a family outing to a local amusement park or water park. • Watch sports games or videos of active events with your child and talk about the benefits of physical activity for overall health. • Encourage your child to invite friends over for a game of soccer, basketball or other team sport to promote socialization and physical activity. By incorporating these fun examples into your family’s routine, you can improve your capabilities for supporting your child’s physical activity and make healthy habits a fun and rewarding part of your family’s lifestyle.</td>
</tr>
</tbody>
</table>

*aM-PAC: Multi-Process Action Control.*

**Phase 2: Recommendations for the Future Use of ChatGPT**

**Overview**

On the basis of our experience in developing the above content and based on previous research [21,25,47], we have compiled the following recommendations for using ChatGPT for similar purposes. First, a 6-step recommendation was proposed to help make the process of using ChatGPT more efficient for future research. These steps included defining the target behavior for the intervention, applying behavior change theory, creating the intervention structure, inputting this information into ChatGPT, refining the output, and customizing the content to be delivered to the target population. Second, we discussed ethical considerations for the use of ChatGPT in academic research. Lastly, we discussed transparency with using ChatGPT in academic research.

**Step 1: Determine Target Behavior**

The first step of these guidelines involves determining a target behavior or behaviors for the target population of your intervention. This could be based on previous research for certain demographic groups or based on a gap in the current literature. The selected target behavior defines the purpose of the intervention and the outcomes on which the intervention will be assessed [25]. This is considered an essential first step as it will guide the remaining steps of these recommendations.

**Step 2: Ground in Behavior Change Theory**

The second step recognizes the need to deliver digital health content grounded in behavior change theory. Based on previous literature and considering the target behavior selected in step 1, it is advised to select a health behavior theory to guide the intervention. Thus, constructs of the behavior change theory must be considered when searching for and developing digital health intervention content. Further, other elements of the intervention, such as BCTs, to strengthen the behavior change theory [26] should be considered during this step.
**Step 3: Design Intervention Structure**

Step 3 involves designing the intervention structure. In this step, the length of the intervention and the length and amount of content to be delivered should be considered first. After this information has been determined, it is recommended to consider the medium of delivery of the digital health intervention content. Previous research has shown varying success for both web-based interventions and mobile-based interventions [48,49]. Additionally, there are important considerations for best practices with delivering content through these different mediums, which are explored later in this development process.

It is important to note that this step may involve an additional agenda. Examining previous literature, using participatory action research or co-design principles, or other methods may be necessary to ensure that you are gathering content that will be both engaging to the participants and promoting adherence to the target behavior.

**Step 4: Input Intervention Structure and Behavior Change Constructs Into ChatGPT**

The next step is to input the information gathered from steps 1 to 3 and create varying prompts into ChatGPT. If this is your first time logging into ChatGPT through OpenAI, you will need to create a free account. Once your account has been created, you may type your prompt into the text box at the bottom of the screen. Determining an optimal prompt to input includes considering the target behavior, the proposed structure of the intervention, the behavior change theory and its constructs, and BCTs. Further, it is important to consider the rules in which ChatGPT delivers its output, for instance, whether you would prefer the response to be in paragraph form or bullet points. This step is iterative as you receive responses and continue to modify your prompt until you receive the desired output. Additionally, it has been previously recommended to consider assigning a role and tone for ChatGPT to embody in its response or to provide a similar example, when available [50].

**Step 5: Revise the Output of ChatGPT**

This step involves revising the response received from ChatGPT. There is a possibility that the language model has created errors or has provided incorrect references with their output. We compared the results with previous literature and revise and adapted as necessary to ensure that the most accurate information is being provided. Including information from the previous literature in the next prompt may continue to provide more refined ChatGPT responses.

**Step 6: Customize the Content to be Delivered**

The final step of this framework is to customize the content to meet the needs of your intervention. This involves considering the layout and design of how you will deliver the content on your selected medium from step 3, as well as any images or graphics used to supplement the given content. This step may involve working with an additional team to develop a web-based or mobile-based platform to support the health behavior change intervention. Further, user experience and design should be considered to improve usability and satisfaction of the content [51-54]. Table 2 summarizes the steps of these guidelines and considerations to meet the needs of each step.

By following these guidelines and using ChatGPT to assist in the rapid creation of digital health content, many ethical considerations arise. The first consideration, as highlighted above, is ensuring that the responses from ChatGPT are accurate and validated to be used as health information in a research study. This can be done by referencing previous literature or creating a panel of experts in the field to review the output created by ChatGPT. Further, it is vital to ensure that users engaging with AI-generated content through ChatGPT or other LLMs are adequately informed about its limitations, decision-making capabilities, and the crucial nature of their involvement. Transparent communication and obtaining informed consent are pivotal to respect user autonomy and comprehension. Although ChatGPT demonstrates remarkable efficiency in generating responses to prompts, evaluating its applicability within the intervention’s context remains crucial to ensure substantial value to using ChatGPT.

As ChatGPT inevitably continues to support academic research across disciplines, it is also important to consider how ChatGPT is being cited by those who use it. There has been a variety of techniques used so far, with some authors including ChatGPT as an author [55] and others acknowledging the use of ChatGPT [34] to assist with their manuscript.
Table 2. Proposed recommendations for developing digital health content using ChatGPT and a summary of considerations for using this tool.

<table>
<thead>
<tr>
<th>Step</th>
<th>Task</th>
<th>Consideration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Determine target behavior</td>
<td>• Previous research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Needs of the target population</td>
</tr>
<tr>
<td>2</td>
<td>Ground in behavior change theory</td>
<td>• Stage of readiness of participants</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Needs of the population group</td>
</tr>
<tr>
<td>3</td>
<td>Design the intervention structure</td>
<td>• Web or mobile based</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Length of the intervention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Amount of content to be delivered in each bout (ie, how many words, characters, or pages of content to be delivered)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Use co-design or other frameworks to ensure that the intervention aligns with the needs of the target population</td>
</tr>
<tr>
<td>4</td>
<td>Input intervention structure and behavior change constructs into ChatGPT</td>
<td>• Structure prompt to input into ChatGPT (considering instruction, context, input data, and output indicator)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Iteratively adapt prompts based on desired output</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Order in which relevant information relating to each construct is delivered, if not predefined by the literature</td>
</tr>
<tr>
<td>5</td>
<td>Revise the output of ChatGPT</td>
<td>• Refer outputs to previous literature to ensure accuracy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Confirm whether references used by ChatGPT are accurate</td>
</tr>
<tr>
<td>6</td>
<td>Customize the content to be delivered</td>
<td>• Layout and design of content</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Images or graphics to supplement text output</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• User experience and design of the intervention platform</td>
</tr>
</tbody>
</table>

Discussion

Principal Findings

The primary objective of this study was to explore the feasibility of using ChatGPT to develop content for a mobile-based JITAI to promote parental support for their children’s PA. The secondary objective was to propose recommendations for using ChatGPT for future work in this area. To our knowledge, the process of using ChatGPT to develop health intervention content has not yet been documented, so we considered the key components required to develop effective behavior change interventions. We found that using ChatGPT was overall acceptable for this case study. However, a human check by researchers in the field is imperative to ensure the relevance and accuracy of the output provided. The use of ChatGPT and similar LLMs is rapidly evolving, and as such, these proposed recommendations are highly dynamic to the developing nature of these technologies.

This study has several implications for researchers using ChatGPT when developing mHealth app content. First, ChatGPT can help researchers improve the efficiency of creating digital health content for various tailored lessons. Previously, it was determined that ChatGPT can expedite the research process by allowing researchers to focus on steps of the research design process that require more human input, for example, focusing on the experimental design [54,57]. The improvement and versatility of text generation, knowledge translation, and literature review have been documented in various studies that have used ChatGPT in health care education [58]. As seen in this study, ChatGPT can help create various versions of content (varying in writing styles and tones) using a series of different prompts. Further, coupled with the efficiency of developing intervention content, this study has highlighted the ability to efficiently create a variety of tailored content specific to PA messaging. The need for more variety and content options has been previously stated as a limitation in previous studies that did not use ChatGPT for the creation of content [28]. Overall, this study highlights one use case that benefited from the use of ChatGPT to rapidly create digital health content. As ChatGPT is in its infancy, we expect it to evolve quickly [58].

Second, this study highlights the current limitations of using ChatGPT for creating mHealth behavior interventions. Although ChatGPT has great potential to improve the efficiency with which digital health content creation can occur, it is not possible to replicate responses by ChatGPT while using the same prompt [58,59]. This unpredictability poses a significant challenge for health researchers and developers who may require stable and reliable outputs [58]. Because of the probabilistic nature of ChatGPT and similar LLMs, the responses generated from ChatGPT are generated based on a probability distribution, meaning the same response will not be generated [60]. Further, a significant concern is the generation of references by ChatGPT that do not exist or are inaccurate. This lack of interpretability hampers the transparency of mHealth content development, making it difficult for researchers to have a clear understanding of the AI’s decision-making process. Other limitations have been recognized by previous work around ChatGPT. These include limited accuracy, bias and limitations of data, lack of context, and the potential of limited engagement with the content [34]. To mitigate these challenges, we highly recommend a rigorous human fact-checking process, as indicated in our recommendations for mHealth intervention content development.
using ChatGPT, and fine-tuning specific prompts to ensure that the information given by ChatGPT is relevant.

Finally, the integration of ChatGPT with existing mHealth app development tools, such as Pathverse, holds the potential to significantly enhance the efficiency and effectiveness of developing and evaluating JITAI apps. By incorporating ChatGPT’s language generation capabilities into Pathverse, developers can expedite the creation of content-rich JITAI apps. Additionally, reinforcement learning algorithms can play a crucial role in JITAI apps by dynamically adapting the intervention based on real-time data and user feedback [61]. Developers can leverage ChatGPT’s language generation capabilities using its application programming interface to assist with content creation [61]. With the integration of ChatGPT, these algorithms can benefit from the AI-generated content to offer more tailored and contextually relevant interventions. By combining the strengths of reinforcement learning and ChatGPT, JITAI apps can become more adaptive and responsive to individual user’s needs, thereby increasing their effectiveness in promoting behavior change and improving health outcomes.

There are several limitations to this study. First, we used ChatGPT to create content for only 1 JITAI, potentially restricting the generalizability of the study findings. Second, because of ChatGPT’s tendency to provide different responses for the same prompt, it was challenging to accurately characterize the content’s reproducibility and consistency. Lastly, as ChatGPT is rapidly evolving, the use case described in this study may have limited applicability a few years from now. We also want to add that although ChatGPT-3 is currently free to use, it is likely that as it improves, it is likely to come with an associated cost.

**Conclusions**

By using ChatGPT, we were able to expedite the process of creating 13 lessons that were guided by the M-PAC framework, thus highlighting the incredible opportunity ChatGPT presents to rapidly create content for various mHealth JITAI apps. Although we found that ChatGPT was acceptable for this case study, we still encourage the cautious use of ChatGPT and other LLMs in similar contexts. The use of ChatGPT expedited the process of content development to 2 months, the bulk of which was spent on reviewing the content by experts in the field before delivering to population groups. This process was imperative to ensure that accurate and relevant content was being created to be delivered. The results from this study found implications in 3 areas. The first is efficiency in generating a variety of content based on different prompts. Second, this study highlighted the potential limitations of ChatGPT, including the inability to replicate responses from the same prompts and the need for human input to ensure that the output from ChatGPT is accurate. Finally, this case study has highlighted the efficiency of using no-code app builders, such as Pathverse, to disseminate information generated by ChatGPT. It is without a doubt that as ChatGPT and other LLMs continue to improve in sophistication and accuracy, they will continue to integrate into intervention design and other various contexts for researchers. Further research and applications of ChatGPT and the guidelines proposed in this study are imminent in this field as we continue to understand ChatGPT.

**Acknowledgments**

The authors acknowledge that ChatGPT was used to generate results for this study. For a summary of the ChatGPT conversations, see Multimedia Appendix 1.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1
ChatGPT Transcript.

[PDF File (Adobe PDF File), 172 KB - mededu_v101e51426_app1.pdf ]

**References**


56. No authors listed. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023;613(7945):612 [FREE Full text] [doi: 10.1038/d41586-023-00191-1] [Medline: 36694020]


60. Why ChatGPT and other LLMs generate different answers to same questions. Ayfie. 2023. URL: https://blog.ayfie.com/why-chatgpt-and-other-llms-generate-different-answers-to-same-questions [accessed 2023-12-14]


Abbreviations

AI: artificial intelligence  
BCT: behavior change technique  
JITAI: just-in-time adaptive intervention  
LLM: large language model  
mHealth: mobile health  
M-PAC: Multi-Process Action Control  
MVPA: moderate-to-vigorous physical activity  
PA: physical activity

©Amanda Willms, Sam Liu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
AI Education for Fourth-Year Medical Students: Two-Year Experience of a Web-Based, Self-Guided Curriculum and Mixed Methods Study

Areeba Abid1, BS; Avinash Murugan2, MD, MBA; Imon Banerjee3, PhD; Saptarshi Purkayastha4, PhD; Hari Trivedi5, MD; Judy Gichoya5, MD

1Emory University School of Medicine, Atlanta, GA, United States
2Yale New Haven Hospital, New Haven, CT, United States
3Mayo Clinic, Phoenix, GA, United States
4Indiana University-Purdue University, Indianapolis, IN, United States
5Department of Radiology, Emory University, Atlanta, GA, United States

Corresponding Author:
Areeba Abid, BS
Emory University School of Medicine
2015 Uppergate Dr
Atlanta, GA, 30307
United States
Phone: 1 (404) 727 4018
Email: areeba.abid@emory.edu

Abstract

Background: Artificial intelligence (AI) and machine learning (ML) are poised to have a substantial impact in the health care space. While a plethora of web-based resources exist to teach programming skills and ML model development, there are few introductory curricula specifically tailored to medical students without a background in data science or programming. Programs that do exist are often restricted to a specific specialty.

Objective: We hypothesized that a 1-month elective for fourth-year medical students, composed of high-quality existing web-based resources and a project-based structure, would empower students to learn about the impact of AI and ML in their chosen specialty and begin contributing to innovation in their field of interest. This study aims to evaluate the success of this elective in improving self-reported confidence scores in AI and ML. The authors also share our curriculum with other educators who may be interested in its adoption.

Methods: This elective was offered in 2 tracks: technical (for students who were already competent programmers) and nontechnical (with no technical prerequisites, focusing on building a conceptual understanding of AI and ML). Students established a conceptual foundation of knowledge using curated web-based resources and relevant research papers, and were then tasked with completing 3 projects in their chosen specialty: a data set analysis, a literature review, and an AI project proposal. The project-based nature of the elective was designed to be self-guided and flexible to each student’s interest area and career goals. Students’ success was measured by self-reported confidence in AI and ML skills in pre and postsurveys. Qualitative feedback on students’ experiences was also collected.

Results: This web-based, self-directed elective was offered on a pass-or-fail basis each month to fourth-year students at Emory University School of Medicine beginning in May 2021. As of June 2022, a total of 19 students had successfully completed the elective, representing a wide range of chosen specialties: diagnostic radiology (n=3), general surgery (n=1), internal medicine (n=5), neurology (n=2), obstetrics and gynecology (n=1), ophthalmology (n=1), orthopedic surgery (n=1), otolaryngology (n=2), pathology (n=2), and pediatrics (n=1). Students’ self-reported confidence scores for AI and ML rose by 66% after this 1-month elective. In qualitative surveys, students overwhelmingly reported enthusiasm and satisfaction with the course and commented that the self-direction and flexibility and the project-based design of the course were essential.

Conclusions: Course participants were successful in diving deep into applications of AI in their widely-ranging specialties, produced substantial project deliverables, and generally reported satisfaction with their elective experience. The authors are
hopeful that a brief, 1-month investment in AI and ML education during medical school will empower this next generation of physicians to pave the way for AI and ML innovation in health care.

(JMIR Med Educ 2024;10:e46500) doi:10.2196/46500

KEYWORDS
medical education; machine learning; artificial intelligence; elective curriculum; medical student; student; students; elective; electives; curricula; curriculum; lesson plan; lesson plans; educators; educator; teacher; teachers; teaching; computer programming; programming; coding; programmer; programmers; self guided; self directed

Introduction

Artificial intelligence (AI) and machine learning (ML) are poised to have a substantial impact in the health care space with many disruptive technologies on the horizon. Innovations in clinical care are increasingly impacted by the development and implementation of AI and ML, and as future clinicians, medical students need to become innovators and active participants in technological changes that will affect how they provide care for their patients. There is much excitement and curiosity among medical students about these technologies [1]. However, few programs exist to deliberately expose future physicians to their role in medicine, let alone to empower students to actively participate in AI and ML innovation [2]. While a plethora of high-quality web-based resources exist to teach programming skills and ML model development, there are few introductory curricula specifically tailored to medical students without a background in data science or programming. Additionally, there is little guidance provided to medical students on where to begin. Some medical societies do have AI outreach activities, but these are limited to trainees within their specialty [3-5].

The authors theorized that a 1-month elective for fourth-year students, composed of existing web-based resources and a project-based structure, would empower students to learn about the impact of AI and ML in their chosen specialty and begin contributing to innovation in their field of interest. The authors also aimed for the elective to be specialty-agnostic and customizable to each student’s career goals. The goal of this senior elective is to demystify AI and ML in health care, enabling students to have informed conversations about these technologies and participate in their clinical advancement. The target participant in the elective is any senior medical student with an interest in AI, with no prerequisites for technical, mathematical, or engineering skills.

In this paper, we evaluate the success of this elective over a 2-year period based on self-reported confidence scores in AI and ML. We also publish our curriculum for other educators who may be interested in its adoption.

Methods

Design

We built our elective following advice on designing medical electives with the principles articulated by Ramalho et al [6], which emphasize that a one-size-fits-all approach is often inadequate and that electives benefit from allowing students to carve their own paths. Creating a medical elective in an overloaded, overworked environment is nontrivial, but prior studies on peer-organized coursework gave us insights into the effectiveness of peer-organized research in building academic confidence, as well as the importance of clearly defined learning objectives [7,8].

Technical and Nontechnical Tracks

Given the wide-ranging skill sets that medical students are equipped with before coming to medical school, this elective was offered in 2 tracks: Technical and Nontechnical. The Technical track was intended for the subset of students who were already competent computer programmers. This course did not aim to teach noncoding students how to code because it was expected that 1 month would not be sufficient time for students to make meaningful progress. Therefore, the Nontechnical track was offered to students with no technical background and focused on building a conceptual understanding of AI. Our goal for the Nontechnical track was to help students without a technical background develop a skill set and vocabulary that would enable them to participate in AI and ML evaluation and implementation processes in future collaborations with technical colleagues.

For both the Technical and Nontechnical tracks, the course was designed to address the following learning objectives:

1. Compare and contrast AI and ML.
2. State and differentiate various ML techniques (supervised/unsupervised, classification/regression, etc).
3. Appreciate the growing impact of ML in medicine, broadly and in the student’s chosen specialty.
4. Develop an intuition of how machines “learn.” Describe how neural networks are structured, trained, and evaluated. Learn vocabulary and concepts used to describe model training (loss functions, gradient descent, and backpropagation).
5. Understand the limitations and pitfalls of ML (reproducibility, interpretability, and bias).
6. Understand what kinds of medical problems can and cannot be solved by ML.
7. Describe issues that may arise in the implementation of an ML algorithm in clinical practice.
8. Discuss ethical issues that concern the use of ML in health care.

Didactic and Project-Based Components

In this self-guided, web-based course, students were referred to existing web-based courses and relevant research papers to supplement these learning objectives (Multimedia Appendix 1 [9-22]) but were expected to guide their own learning beyond this. Students were asked to share and write down their personal
goals at the beginning of the elective to guide their learning. They were also encouraged to spend time after each section on independent research to address lingering questions. The learning objectives and course resources were provided to students on a central document and students were able to follow along at their own pace. Because the course aimed to empower an individual student’s interests and career goals, the elective was designed to establish a baseline level of understanding for all students, while also allowing students the freedom to dive deeper into the areas they were drawn to. Students were supported by the course’s faculty advisor, a physician with substantial leadership and experience in AI and ML research.

Project Deliverables
Students were then tasked with completing at least 1 of the following project-based deliverables, and encouraged to complete others as their interests dictated:

1. Complete a literature review on the state of AI and ML in the student’s chosen specialty.
2. Find and analyze 3 open-source health care data sets, considering strengths, weaknesses, and sources of error and bias.
3. Write a Project Proposal addressing a problem in the student’s chosen specialty that can be solved with AI, with a discussion surrounding the implementation complexities.
4. Technical track only: Train and evaluate a clinical ML algorithm.

Details on these projects are provided in Multimedia Appendix 2 [23].

The full curriculum is hosted on the Emory Health Care Innovations and Translational Informatics Lab GitHub repository [24].

This course was initially designed during the COVID-19 pandemic, and maintained a web-based format throughout the 2 years it has been offered. All recommended resources were freely available to students on the web, although some required institutional access. The students attended weekly web-based laboratory meetings to discuss their progress and to be exposed to more advanced research in AI and ML. Students were also encouraged to identify an additional advisor (beyond the elective director, who they met with once a week) within their chosen specialty, who could provide domain expertise for their projects.

Qualitative Survey Data
Initially, the authors collected feedback from students qualitatively through one-on-one meetings; this feedback was used to improve the format and support structure of the elective. Beginning in October 2021, students were also asked for open-ended feedback on the strengths and weaknesses of the elective through anonymous surveys. They were asked:

- What was the most meaningful project or experience you completed during the elective? Do you intend to continue work on it past the end of the elective?
- Did you gain what you hoped to get out of this elective? Please explain.
- What resources were most useful to you during the elective?
- What could be most improved in the curriculum design of this elective?

Quantitative Survey Data
Beginning in October 2021, quantitative pre and postelective surveys were implemented using Google Forms to assess the effectiveness of the elective format and resources provided. Students were asked to fill out formal surveys to rate their confidence in AI and ML concepts and in technical data science and coding skills.

Before starting the elective, students were asked:

- How familiar are you with AI or ML concepts? (Likert scale, 1-5)
- How would you rate your technical data science or coding experience? (Likert scale, 1-5)

After completing the elective, students were asked:

- Did you choose the Technical or Nontechnical Track?
- After completing this elective, how familiar are you with AI or ML concepts? (Likert scale, 1-5)
- After completing this elective, how would you rate your technical data science or coding experience? (Likert scale, 1-5)

Statistical Analysis
Quantitative and discrete data from self-reported confidence scores was analyzed using the Wilcoxon rank sum test. Qualitative survey responses were reviewed in a descriptive manner rather than undergoing a formal analysis. Responses were manually examined for common themes, trends, and noteworthy insights, but no systematic coding framework was used and representative responses are included in the “Results” section.

Ethical Considerations
This study was deemed exempt from review by Emory University’s institutional review board, under the category “Educational Tests, Surveys, Interviews, Observations.” This is justified based on anonymity and minimal risk to survey participants. All participants were able to opt out of this educational experience and from data collection. Survey data were collected anonymously. Students were not compensated for participation.

Results
Overview
This web-based, self-directed elective was offered on a pass-or-fail basis each month to fourth-year students at Emory University School of Medicine beginning in May 2021. A maximum of 3 students were allowed to enroll each month. As of June 2022, a total of 19 students had signed up and completed the elective. All students successfully met elective requirements and passed the course. The students represented a diverse range of chosen specialties: diagnostic radiology (n=3), general surgery (n=1), internal medicine (n=5), neurology (n=2), obstetrics and gynecology (n=1), ophthalmology (n=1),...
orthopedic surgery (n=1), otolaryngology (n=2), pathology (n=2), and pediatrics (n=1).

Given the limited time and open-ended nature of the course, students elected to spend varying amounts of time on each of the project components based on their interests and were not required to complete all 3 projects as long as they produced at least 1 significant deliverable. The vast majority of students (17 out of 19 students) chose the Nontechnical track. Most students (11/19, 58%) chose to focus their efforts on 2 of the 3 projects; 8 (42%) completed all 3 projects, and 1 (5%) submitted only a project proposal. Since the elective was intended to be flexible to students’ interests, students were evaluated on a pass-or-fail basis based on demonstrated effort as determined by the faculty advisor, rather than strict adherence to project deliverables. All students received a passing grade. Project proposals submitted by students were wide-ranging, including AI applications such as “Smartphone Detection of Anterior Uveitis,” “Predicting Postpartum Hemorrhage,” “Image Enhancement in Video Laryngoscopy,” and “Audiometry for Pediatric Heart Murmur Screening.” Four (25%) students indicated that they intended to continue working on their projects beyond the end of the elective.

Qualitative Survey Results

Qualitative feedback collected from students before October 2021 (n=4) indicated that students wanted more support and guidance in their field of interest; given this feedback, the authors created more structure for the elective and encouraged students to find an additional specialty-specific mentor who could contribute domain expertise.

Students were asked if they gained what they hoped for from their elective experience. Students who sought a basic conceptual understanding reported satisfaction, but some reported an unmet desire for a deeper technical understanding:

- “I wanted to learn more generally how AI/ML can be used and is being used in medicine. I definitely achieved this goal.”
- “I feel that I learned AI/ML fundamentals, am now able to better read and understand AI/ML medical literature, and have thought through the essential design elements of an AI/ML proposal.”
- “I learned about the clinical applications of ML and how it is used to help rather than replace radiologists. I also have learned that the technology is advanced, but the application is still early in medicine.”
- “I found the course very valuable as an introduction to what ML is and how it is used. However, I had hoped to gain more insight into what research is being conducted in ML from a technical perspective and what these advances may mean from a translational perspective.”

Students were also asked what aspects of the course were most beneficial. Four students commented that the self-directed and flexible nature of the course was essential. Two students commented that the project proposal was the most essential element. Five (26%) students reported that they intended to continue working on their projects after the end of the elective month.

When asked for constructive feedback, 2 students commented that they desired more concrete guidance on the projects. Some students felt strained to finish the project proposal within 1 month, with one commenting that students should not expect to finish the proposal in 1 month, and 2 recommending future students pick a project as early as possible, rather than waiting until after the literature review and data set project.

Quantitative Survey Results

After October 2021, students were asked to fill out formal surveys collecting feedback and self-reported confidence in skills gained during the elective. Fifteen students filled out the preintervention survey, and 12 students completed the postintervention survey. These results are shown in Table 1.

Table 1. Pre- and postintervention confidence scores in AI or ML concepts and technical skills.

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>Median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>“On a scale of 1-5, how well do you understand AI or ML concepts?”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preintervention (n=15)</td>
<td>2.5 (1.3)</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Postintervention (n=12)</td>
<td>4.1 (0.7)</td>
<td>4 (3)</td>
</tr>
<tr>
<td><strong>“On a scale of 1-5, rate your technical data science skills”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preintervention (n=15)</td>
<td>2.6 (1.4)</td>
<td>3 (0.25)</td>
</tr>
<tr>
<td>Postintervention (n=12)</td>
<td>1.9 (1.3)</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

a AI: artificial intelligence.
b ML: machine learning.
c Relative difference is 66% and Wilcoxon rank sum P value is .003.
d Relative difference is –26% and Wilcoxon rank sum P value is .20.
Discussion

Principal Results

Students who participated in this elective were successful in diving deep into the potential of AI and ML in their area of interest and generally reported satisfaction with their elective experience. Students were asked to quantitatively rate their familiarity with both AI and ML concepts and coding or data science; the self-reported confidence scores for AI and ML rose by 66%, and these results were found to be statistically significant when analyzed by the Wilcoxon rank sum test. This exposure to AI and ML is a substantial improvement from the status quo, in which most medical students receive little to no exposure during the course of their training; in 1 study from 2022, 66.5% of students reported 0 hours of AI or ML teaching, and 43.4% had never heard the term “machine learning” [25]. Previous literature includes effective AI curricula developed for other types of health care trainees, such as radiology residents, but there is little to no literature on curricula evaluated for a fourth-year medical student audience as described in this paper [26,27].

Self-reported confidence in technical skills (coding and data science) fell by 26%, although this result was not found to be statistically significant. The authors attribute these results to an initial overconfidence prior to the elective, followed by an increased awareness of the technical complexity of model development after the elective.

Because this was a self-guided elective, student output varied with each student’s level of motivation and goals prior to entering the elective. Students who had defined a specific area of interest tended to benefit more from their experience than students who came in with no clear goals set. This course could be improved by providing further assistance early on in helping students to finalize a project area early so that they feel less strained by time toward the end of the month.

Students produced a wide range of deliverables in their chosen specialty. Since most fourth-year students have chosen their specialty and have established connections with faculty in their field, the self-guided nature of the course allowed flexibility for students to seek out appropriate mentors and propose reasonable projects in their areas of interest.

Limitations and Future Directions

Limitations of this study include the small number of participants, especially in the Technical track, restricting the generalizability of this study. Only 2 (11%) students chose the Technical track, so there is insufficient data to evaluate this curriculum; this was likely due to the requirement that students interested in the Technical track have in-depth coding experience and receive approval from the course director to ensure a high likelihood of success. However, the authors recommend screening applicants to make sure that they do in fact possess the required level of comfort in coding before attempting to develop an ML model, as we observed a tendency for students to underestimate the complexity of this task. Based on qualitative observations that students spent more time than expected preparing data for training, the authors suggest providing select, cleaned data sets for students in the Technical track, allowing them to focus on model building, training, and testing.

Another substantial limitation is that assessments relied only on students’ self-reported confidence, which has been shown to be a flawed metric [28]. Further studies would benefit from a refined objective assessment tool of students’ competencies, as well as replication of this study at other medical schools.

Since launching this fourth-year elective, we have also adapted this curriculum to a shorter elective targeting second-year medical students and were invited to participate in a National Academies forum on AI for Health Profession Education to disseminate this curriculum to other learners [29].

Conclusions

Overall, in the 2 years since launching the elective at Emory University School of Medicine, the authors have already seen substantial excitement and appreciation from senior medical students, with continued excitement in the elective’s third year. Most students entered the elective with minimal previous experience in AI and ML and were successful in completing self-guided research and proposing creative and realistic AI and ML projects. The authors are hopeful that a brief, 1-month investment in AI and ML education during medical school can lay the groundwork for these future physicians to continue to engage with AI and ML research and empower this next generation of physicians to pave the way for AI and ML innovation in health care.

Acknowledgments

This study would not have been possible without the support of Emory University School of Medicine. The authors are grateful to Meredith Greer for her guidance in curricular development.

Data Availability

The data sets generated or analyzed during this study are not publicly available due to ensure participant confidentiality and privacy in compliance with the institutional review board exemption status, but are available from the corresponding author on reasonable request.
Authors' Contributions

AA and JG contributed to the conceptualization, investigation, and methodology; analysis of results; and the writing of the manuscript. AM contributed to the conceptualization and design of the course, along with the review and editing of the manuscript. IB, SP, and HT contributed to the administration of the elective and review and editing of the manuscript.

Conflicts of Interest

JG is a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program and declares support from Radiological Society of North America Health Disparities grant (#EIHDD204), Lacuna Fund (#67), Gordon and Betty Moore Foundation, and National Institutes of Health (National Institute of Biomedical Imaging and Bioengineering) Medical Imaging and Data Resource Center grant (contracts 75N92020C00008 and 75N92020C00021). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Multimedia Appendix 1
Learning objectives and corresponding curated resources.

[DOCX File, 17 KB - mededu_v10i1e46500_app1.docx ]

Multimedia Appendix 2
Project components and deliverables.

[DOCX File, 16 KB - mededu_v10i1e46500_app2.docx ]

References


Abbreviations

AI: artificial intelligence
ML: machine learning
Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard

Faiza Farhat¹, PhD; Beenish Moalla Chaudhry², PhD; Mohammad Nadeem³, PhD; Shahab Saquib Sohail⁴, PhD; Dag Øivind Madsen⁵, PhD

¹Department of Zoology, Aligarh Muslim University, Aligarh, India
²School of Computing and Informatics, The University of Louisiana, Lafayette, LA, United States
³Department of Computer Science, Aligarh Muslim University, Aligarh, India
⁴School of Computing Science and Engineering, VIT Bhopal University, Sehore, India
⁵School of Business, University of South-Eastern Norway, Hønefoss, Norway

Abstract

Background: Large language models (LLMs) have revolutionized natural language processing with their ability to generate human-like text through extensive training on large data sets. These models, including Generative Pre-trained Transformers (GPT)-3.5 (OpenAI), GPT-4 (OpenAI), and Bard (Google LLC), find applications beyond natural language processing, attracting interest from academia and industry. Students are actively leveraging LLMs to enhance learning experiences and prepare for high-stakes exams, such as the National Eligibility cum Entrance Test (NEET) in India.

Objective: This comparative analysis aims to evaluate the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions.

Methods: In this paper, we evaluated the performance of the 3 mainstream LLMs, namely GPT-3.5, GPT-4, and Google Bard, in answering questions related to the NEET-2023 exam. The questions of the NEET were provided to these artificial intelligence models, and the responses were recorded and compared against the correct answers from the official answer key. Consensus was used to evaluate the performance of all 3 models.

Results: It was evident that GPT-4 passed the entrance test with flying colors (300/700, 42.9%), showcasing exceptional performance. On the other hand, GPT-3.5 managed to meet the qualifying criteria, but with a substantially lower score (145/700, 20.7%). However, Bard (115/700, 16.4%) failed to meet the qualifying criteria and did not pass the test. GPT-4 demonstrated consistent superiority over Bard and GPT-3.5 in all 3 subjects. Specifically, GPT-4 achieved accuracy rates of 73% (29/40) in physics, 44% (16/36) in chemistry, and 51% (50/99) in biology. Conversely, GPT-3.5 attained an accuracy rate of 45% (18/40) in physics, 33% (13/26) in chemistry, and 34% (34/99) in biology. The accuracy consensus metric showed that the matching responses between GPT-4 and Bard, as well as GPT-4 and GPT-3.5, had higher incidences of being correct, at 0.56 and 0.57, respectively, compared to the matching responses between Bard and GPT-3.5, which stood at 0.42. When all 3 models were considered together, their matching responses reached the highest accuracy consensus of 0.59.

Conclusions: The study’s findings provide valuable insights into the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions. GPT-4 emerged as the most accurate model, highlighting its potential for educational applications. Cross-checking responses across models may result in confusion as the compared models (as duos or a trio) tend to agree on only a little over half of the correct responses. Using GPT-4 as one of the compared models will result in higher accuracy consensus. The results underscore the suitability of LLMs for high-stakes exams and their positive impact on education. Additionally, the
study establishes a benchmark for evaluating and enhancing LLMs’ performance in educational tasks, prompting responsible and informed use of these models in diverse learning environments.

(JMIR Med Educ 2024;10:e51523) doi:10.2196/51523

KEYWORDS
accuracy; AI model; artificial intelligence; Bard; ChatGPT; educational task; GPT-4; Generative Pre-trained Transformers; large language models; medical education, medical exam; natural language processing; performance; premedical exams; suitability

Introduction

Large language models (LLMs) are potent natural language processing tools, excelling in a range of artificial intelligence (AI) tasks, from news writing to product descriptions. They have garnered widespread attention across academia and industry [1,2], going beyond the scope of natural language processing into tasks related to health care [3], neuroscience [4], philosophy [5], marketing and finance [6,7], sociology [8], education, and others [9,10]. The development of LLMs and chatbots is experiencing an upsurge, with established companies and emerging start-ups actively engaged in their creation [11], catering to general or specific purposes [12]. Prominent examples include Generative Pre-trained Transformers (GPT)-3.5 (OpenAI), GPT-4 (OpenAI), and Bard (Google LLC) [13,14]. Other notable examples are BlenderBot, Galactica, LLaMA (FAIR) [15], Alpaca (Stanford), BloombergGPT [16], Chinchilla (DeepMind), and PaLM [17], heralding the emergence of even more chatbots in the future [12].

The public release of ChatGPT in November 2022 and Bard in March 2023 has garnered significant attention due to their general purpose and flexible nature. ChatGPT [18], built on the GPT-3.5 architecture, has become popular for its remarkable ability to generate coherent and human-like responses. GPT-4.0 represents the latest iteration, incorporating enhanced language generation and improved multiturn conversation handling. Both GPT-3.5 and GPT-4.0 have been specifically trained to interact with users in a conversational manner, maintaining context, handling follow-up questions, and even correcting themselves. Bard, on the other hand, leverages Google’s LaMDA [19], enabling it to handle a diverse range of language-related tasks and provide in-depth information.

In educational settings, students are using LLMs such as Bard, GPT-3.5, and GPT-4 to enrich their daily learning experiences [20,21]. They aid students in test preparation, offer research assistance, and contribute to their overall performance improvement and knowledge acquisition [22]. It has been observed that LLMs, despite their impressive performance, can sometimes generate text that includes fabricated or incorrect information [13,23]. Consequently, researchers have directed their attention toward investigating the test-taking capabilities of different LLMs. Numerous research studies have delved into the assessment of GPT-3.5’s efficacy in multiple-choice exams in higher education domains [24]. Some investigations have specifically focused on ChatGPT’s test-taking performance in diverse professional fields, including business [25], accounting [26], law [27], and medicine [28]. In the medical realm, authors in Bommineni et al [29] examined its competence in tackling the Medical College Admissions Test, which serves as a prerequisite for admission to most medical schools in the United States. In Gilson et al [30] and Kung et al [31], authors have scrutinized ChatGPT’s aptitude in the United States Medical Licensing Examination (USMLE), while Teebagy et al [32] conducted a comparative study of GPT-3.5 and GPT-4’s performance in the Ophthalmic Knowledge Assessment Program exam. Additionally, Ali et al [33] undertook a comparison of GPT-3.5, GPT-4, and Google Bard, using questions specifically prepared for neurosurgery oral board examinations. Similarly, Zhu et al [28] investigated ChatGPT’s performance in several medical topics, namely, the American Heart Association, advanced cardiovascular life support, and basic life support exams.

Despite the successful integration of LLMs in educational environments, a crucial question remains: can LLMs provide the necessary accuracy and reliability required for critical assessments? The published studies predominantly focus on specialized fields within medicine, with few investigations addressing the effectiveness of AI tools for medical school entrance examinations [29]. Additionally, such comparisons made in the literature typically revolve around the performance of a solitary LLM against human abilities [24,34], with limited exploration of how they compare against other LLMs or baseline models, which could provide valuable insights into the strengths and weaknesses of different LLMs. Our primary objective is to bridge this knowledge gap by undertaking a comparative analysis of 3 notable chatbots: GPT-3.5, GPT-4, and Bard, for a standardized medical school exam known as the National Eligibility cum Entrance Test (NEET).

NEET [35] is a competitive entrance exam in India for Bachelor of Medicine and Bachelor of Dental Surgery programs in both government and private colleges. Introduced in 2013 by the Medical Council of India, NEET replaced various state-level and institution-specific tests to standardize medical admissions. Since 2019, the National Testing Agency (NTA) has been responsible for conducting and supervising the NEET. The exam comprises a total of 200 multiple-choice questions aimed at testing knowledge, understanding, and aptitude in 4 subjects: physics, chemistry, botany, and zoology. Candidates can only attempt a maximum of 45 questions per subject, for a total of 180 out of 200 questions. Correct answers are awarded 4 points, while each incorrect response leads to a 1-point deduction. Candidates are allotted 3 hours to complete the examination. To qualify for admission to a medical school, candidates must obtain a minimum or cutoff score, which can change year by year. The cutoff score for NEET-2023 was 137 out of 720. In 2023, over 2.03 million students took the NEET exam [24], a number that has been rising annually by 10% to 16.5%, highlighting the exam’s widespread popularity and importance.
Among the 1.15 million candidates who qualified in 2023, only 2 scored full marks (720/720), only 1 scored 716 out of 720, a total of 17 scored 715 out of 720, and 6 scored 711 out of 720 [36]. NEET’s rigorous nature, coupled with its widespread adoption, underscores its importance as the primary evaluation tool for determining students’ knowledge, aptitude, and readiness for pursuing medical and dental education at the undergraduate level [35].

In this investigation, to evaluate the performance of the 3 mainstream LLMs, namely GPT-3.5, GPT-4, and Google Bard, in answering questions related to the NEET 2023 exam, we used rigorous statistical analyses. We scrutinized each model’s performance across 3 pivotal frameworks: overall comparison, subject-level comparison, and topic-level comparison. The outcomes of this study can help premed students make informed decisions about incorporating LLMs into their test preparation strategies. To the best of our knowledge, this marks the first endeavor to undertake such a study.

**Methods**

**Question Set Selection and Preparation**

In this paper, we tested the performance of the 3 LLMs on NEET-2023, which was obtained as a portable document file. Although the exam consists of 200 questions, due to the presence of illustrations and diagrams, it was not possible to process all the questions. As a result, we excluded questions with illustrations, resulting in a set of 175 questions for this study. This sample size is large enough to statistically justify each model’s performance on the entire exam, with a 95% CI and a 5% margin of error. The selected questions were then manually presented to Bard, GPT-3.5, and GPT-4, and the responses were documented in Excel (Microsoft Corporation).

**Data Analysis**

We compared responses generated by each model against the correct answers from the official answer key on the NEET website. Based on this comparison, the responses were either marked as correct (1) or incorrect (0).

**Prediction Performance**

Excel’s built-in functionalities were then used to generate the following comparison metrics to assess predictive performance of the LLMs:

1. Accuracy is defined as the percentage of correct responses obtained by a model. In the context of this research, accuracy was obtained using the formula:
   \[
   \text{Accuracy} = \frac{\text{Correct Responses}}{\text{Total Responses}}
   \]

2. Accuracy consensus is defined as the ratio between correct answers upon which the compared models agree to all the answers (correct and incorrect) upon which the compared models agree. The formula is
   \[
   \text{Accuracy consensus} = \frac{\text{Correct Responses}}{\text{Total Consensus}}
   \]

**Results**

**Prediction Performance**

Next, we calculated the overall, subject-level, and topic-level percentage scores for each LLM following the NTA’s scoring rules. Each correct answer was awarded 4 points, while each incorrect answer resulted in a deduction of 1 point. We merged zoology and botany into a single biology category, as the topic-level analysis included questions from both fields. The overall score percentage for each model was determined by dividing the total points scored by the maximum possible points, which was 700. Subject-level percentages were derived by dividing each model’s total points by the maximum points available in that subject. Similarly, topic-level percentages were calculated by dividing the total points scored in each topic by the maximum points available for that topic, which varied across different topics.

**Scoring Performance**

The results demonstrated that GPT-4 had higher accuracy and consensus compared to GPT-3.5 and Bard. It also consistently outperformed the other models across subjects and topics. GPT-3.5 and Bard showed variations in their performances, with specific strengths in certain subjects and topics.

**Overall Accuracy**

The overall accuracy rates of the models were as follows:

1. GPT-4 achieved the highest accuracy rate of approximately 54.3% by correctly identifying 95 out of 175 responses.
2. GPT-3.5 demonstrated an accuracy of 36.7%, with 64 out of 175 correct responses.
3. Bard achieved the lowest accuracy of approximately 33.1%, based on 58 out of 175 correct answers.

**Subject-Level Accuracy**

Table 1 presents the number of correct responses obtained by each model in each of the 3 subject areas covered by NEET. It was evident that GPT-4 is consistently more accurate than both Bard and GPT-3.5 in all 3 subjects. For each subject, the number of correct responses obtained by GPT-3.5 and Bard differed by ±3, indicating relatively similar subject-level accuracy rates. On the other hand, GPT-4 was substantially more accurate than the other models, generating 4 to 16 more correct answers per subject. In physics, GPT-4 achieved 73% (29/40) accuracy, followed by GPT-3.5 with 45% (18/40), and Bard with 38% (15/40). Similarly, in chemistry, GPT-4’s accuracy rate was 44% (16/36), while GPT-3.5 and Bard achieved an accuracy rate of 33% (12/36). Shifting to biology, GPT-4 maintained its lead with 51% (50/99) accuracy, followed by GPT-3.5 with 34% (34/99), and then Bard with 31% (31/99).
Table 1. Number of correct responses (n) and accuracy rates in each subject per model.

<table>
<thead>
<tr>
<th>Subject</th>
<th>GPT\textsuperscript{a}-4, n (%)</th>
<th>GPT-3.5, n (%)</th>
<th>Bard, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology (n=99)</td>
<td>50 (51)</td>
<td>34 (34)</td>
<td>31 (31)</td>
</tr>
<tr>
<td>Chemistry (n=36)</td>
<td>16 (44)</td>
<td>12 (33)</td>
<td>12 (33)</td>
</tr>
<tr>
<td>Physics (n=40)</td>
<td>29 (73)</td>
<td>18 (45)</td>
<td>15 (38)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}GPT: Generative Pre-trained Transformers.

**Topic-Level Accuracy**

Table 2 displays the number of correct responses obtained from each model on various topics. GPT-4 was the most accurate in 9 (50%) out of 18 topics. Moreover, for at least half (2-4) of the topics in each subject, GPT-4 demonstrated the highest accuracy. GPT-3.5 was the most accurate (8/15, 53%) in inorganic chemistry. In addition, it was more accurate than Bard in 7 topics across the 3 subjects. However, it had a 0% accuracy in population and ecology (biology) and simple harmonic motion and waves (physics). Bard was the most accurate in the topics on plant kingdom and ecosystem and environment issues. Furthermore, it was more accurate than GPT-3.5 in 5 topics across all 3 subjects. However, it has a 0% accuracy for 2 physics topics, namely modern physics and electronics and optics. GPT-4 and GPT-3.5 had similar accuracies in 1 physics topic (modern physics and electronics: 2/4, 50%) and 2 biology topics (cell biology and genetics: 7/16, 44%; and ecosystem and environmental issues: 2/5, 40%). GPT-4 and Bard are 100% accurate in the topics on simple harmonic motion and waves. All 3 models were at the same level of accuracy in the topics on biomolecules and heat and thermodynamics.

In a nutshell, GPT-4 had a higher accuracy across a wide range of topics (15/18, 83%), while GPT-3.5’s and Bard’s accuracies were well below GPT-4’s. Moreover, they showed variations in their accuracies across topics.

Table 2. Number of correct responses for each topic per model.

<table>
<thead>
<tr>
<th>Topic</th>
<th>GPT\textsuperscript{a}-4, n (%)</th>
<th>GPT-3.5, n (%)</th>
<th>Bard, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biotechnology (n=11)</td>
<td>7 (64)\textsuperscript{b}</td>
<td>6 (55)</td>
<td>4 (36)</td>
</tr>
<tr>
<td>Evolution and health (n=9)</td>
<td>7 (78)\textsuperscript{b}</td>
<td>4 (44)</td>
<td>2 (22)</td>
</tr>
<tr>
<td>Population and ecology (n=6)</td>
<td>1 (17)\textsuperscript{b}</td>
<td>0 (0)</td>
<td>1 (17)\textsuperscript{b}</td>
</tr>
<tr>
<td>Biomolecules (n=3)</td>
<td>1 (33)</td>
<td>1 (33)</td>
<td>1 (33)</td>
</tr>
<tr>
<td>Cell biology and genetics (n=16)</td>
<td>7 (44)\textsuperscript{b}</td>
<td>7 (44)\textsuperscript{b}</td>
<td>3 (19)</td>
</tr>
<tr>
<td>Ecosystem and environmental issues (n=5)</td>
<td>2 (40)</td>
<td>2 (40)</td>
<td>3 (60)\textsuperscript{b}</td>
</tr>
<tr>
<td>Plant kingdom (n=25)</td>
<td>8 (32)</td>
<td>6 (24)</td>
<td>11 (44)\textsuperscript{b}</td>
</tr>
<tr>
<td>Animal kingdom (n=24)</td>
<td>17 (71)\textsuperscript{b}</td>
<td>8 (33)</td>
<td>6 (25)</td>
</tr>
<tr>
<td>Physical chemistry (n=12)</td>
<td>6 (50)\textsuperscript{b}</td>
<td>3 (25)</td>
<td>4 (33)</td>
</tr>
<tr>
<td>Organic chemistry (n=9)</td>
<td>3 (33)\textsuperscript{b}</td>
<td>1 (11)</td>
<td>2 (22)</td>
</tr>
<tr>
<td>Inorganic chemistry (n=15)</td>
<td>7 (47)</td>
<td>8 (53)\textsuperscript{b}</td>
<td>6 (40)</td>
</tr>
<tr>
<td>Mechanics (n=12)</td>
<td>8 (67)\textsuperscript{b}</td>
<td>6 (50)</td>
<td>6 (50)</td>
</tr>
<tr>
<td>Heat and thermodynamics (n=3)</td>
<td>1 (33)</td>
<td>1 (33)</td>
<td>1 (33)</td>
</tr>
<tr>
<td>Electrostatics and electricity (n=11)</td>
<td>10 (91)\textsuperscript{b}</td>
<td>5 (45)</td>
<td>6 (55)</td>
</tr>
<tr>
<td>Optics (n=3)</td>
<td>3 (100)\textsuperscript{b}</td>
<td>2 (67)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Simple harmonic motion and waves (n=1)</td>
<td>1 (100)\textsuperscript{b}</td>
<td>0 (0)</td>
<td>1 (100)\textsuperscript{b}</td>
</tr>
<tr>
<td>Magnetism (n=6)</td>
<td>4 (67)\textsuperscript{b}</td>
<td>2 (33)</td>
<td>1 (17)</td>
</tr>
<tr>
<td>Modern physics and electronics (n=4)</td>
<td>2 (50)\textsuperscript{b}</td>
<td>2 (50)\textsuperscript{b}</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}GPT: Generative Pre-trained Transformers.

\textsuperscript{b}Highest accuracy within a topic.
Accuracy Consensus

Overall Accuracy Consensus

The accuracy consensus for the pairs were approximately as follows:

1. Bard and GPT-3.5 were correct on 29 out of 69 matching responses, giving the pair an accuracy consensus of 0.42 and an accuracy of 29 (16.6%) out of 175.

2. Bard and GPT-4 were correct on 42 out of 75 matching responses, resulting in an accuracy consensus of 0.56 and an accuracy of 42 (24%) out of 175.

3. GPT-3.5 and GPT-4 were correct on 45 out of 79 matching responses, giving the pair an accuracy consensus of 0.57 and an accuracy of 45 (25.7%) out of 175.

4. All 3 models were correct on 29 out of 49 matched responses. The accuracy consensus of the trio was approximately 0.59 and an accuracy of 29 (16.6%) out of 175.

This ascending trend in accuracy consensus indicated that GPT-4 enhanced the agreement on correct responses, especially when used in conjunction with either Bard or GPT-3.5. The best accuracy consensus and accuracy were obtained when GPT-3.5 and GPT-4 were considered together. Moreover, the collective intelligence of these models was as good as the weakest duo, that is, Bard and GPT-3.5 combined.

Subject-Level Accuracy Consensus

Table 3 shows the total number of correct matching responses and accuracy consensus at the subject level for each model.

Table 3. Subject-level total correct matching responses and accuracy consensus across compared models.

<table>
<thead>
<tr>
<th>Subject</th>
<th>GPT-3.5 vs Bard</th>
<th>Bard vs GPT-4</th>
<th>GPT-3.5 vs GPT-4</th>
<th>Bard, GPT-3.5, and GPT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total correct matching responses, n</td>
<td>Accuracy consensus</td>
<td>Total correct matching responses, n</td>
<td>Accuracy consensus</td>
</tr>
<tr>
<td>Biology</td>
<td>17</td>
<td>0.4</td>
<td>22</td>
<td>0.46</td>
</tr>
<tr>
<td>Chemistry</td>
<td>4</td>
<td>0.31</td>
<td>7</td>
<td>0.50</td>
</tr>
<tr>
<td>Physics</td>
<td>8</td>
<td>0.58</td>
<td>13</td>
<td>1.00</td>
</tr>
</tbody>
</table>

⁵GPT: Generative Pre-trained Transformers.
⁴Highest accuracy within a subject.

The subject-level accuracy consensus revealed following insights.

For biology, the highest accuracy consensus was observed between GPT-3.5 and GPT-4 (n=23, ratio of 0.48), indicating GPT-4’s superior performance. This duo also produced the highest accuracy, that is, 23 (23%) out of 99. Even though the accuracy consensus of the trio was the highest, it did not correspond to the highest accuracy (17/99, 17%).

For chemistry, both comparisons involving GPT-4 (Bard vs GPT-4 and GPT-3.5 vs GPT-4) yielded a higher accuracy consensus ratio of 0.50. However, the duo of GPT-3.5 and GPT-4 resulted in highest accuracy, that is, 8 (22%) out of 36.

For physics, Bard versus GPT-4 and the collective comparison of all models achieved a perfect accuracy consensus of 1.00 and an accuracy of 13 (32%) out of 40. However, the highest accuracy (14/40, 35%) was shown by GPT-3.5 versus GPT-4, with comparable accuracy consensus of 0.93.

These points demonstrate GPT-4’s dominance across subjects, with physics showcasing the highest consensus scores. This suggests that when GPT-4 is used in tandem with any other model, the duo or trio will corroborate each other’s responses more than when Bard and GPT-3.5 are considered together.

Topic-Level Accuracy Consensus

Table 4 shows the total number of correct matching responses and accuracy consensus at the topic level for each model.

The following observations can be made about data presented in Table 4.

GPT-3.5 versus GPT-4 demonstrated the highest accuracy consensus and number of correct matching responses in 11 (61%) out of 18 topics. This trend was followed by the Bard versus GPT-4 duo, which showed the highest number of accurate responses and accuracy consensus in 7 (39%) out of 18 topics.

“Biomolecules,” “heat and thermodynamics,” “optics,” and “simple harmonic motion and waves” had low or zero accuracy consensus for all or most comparisons.

Hence, the combined intelligence of the models cannot help with the preparation of all the topics, if the goal is to seek consensus or confirmation of responses across models.
Table 4. Topic-level correct matching responses and accuracy consensus across compared models.

<table>
<thead>
<tr>
<th>Topic</th>
<th>GPT\textsuperscript{a}-3.5 vs Bard</th>
<th>Bard vs GPT-4</th>
<th>GPT-3.5 vs GPT-4</th>
<th>Bard, GPT-3.5, and GPT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total correct matching responses, n</td>
<td>Accuracy consensus</td>
<td>Total correct matching responses, n</td>
<td>Accuracy consensus</td>
</tr>
<tr>
<td>Biotechnology</td>
<td>3</td>
<td>0.75</td>
<td>3</td>
<td>0.60</td>
</tr>
<tr>
<td>Evolution and health</td>
<td>3</td>
<td>0.75\textsuperscript{b}</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>Population and ecology</td>
<td>2</td>
<td>0.67</td>
<td>2</td>
<td>0.67</td>
</tr>
<tr>
<td>Biomolecules</td>
<td>0</td>
<td>N/A\textsuperscript{c}</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>Cell biology and genetics</td>
<td>3</td>
<td>0.30</td>
<td>3</td>
<td>0.43\textsuperscript{b}</td>
</tr>
<tr>
<td>Ecosystem and environmental issues</td>
<td>1</td>
<td>0.33</td>
<td>2</td>
<td>0.67\textsuperscript{b}</td>
</tr>
<tr>
<td>Plant kingdom</td>
<td>2</td>
<td>0.22</td>
<td>4</td>
<td>0.31\textsuperscript{b}</td>
</tr>
<tr>
<td>Animal kingdom</td>
<td>3</td>
<td>0.38</td>
<td>5</td>
<td>0.50\textsuperscript{b}</td>
</tr>
<tr>
<td>Physical chemistry</td>
<td>2</td>
<td>0.67</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>Organic chemistry</td>
<td>1</td>
<td>0.50</td>
<td>3</td>
<td>0.75\textsuperscript{b}</td>
</tr>
<tr>
<td>Inorganic chemistry</td>
<td>1</td>
<td>0.13</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>Mechanics</td>
<td>2</td>
<td>0.50</td>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>Heat and thermodynamics</td>
<td>0</td>
<td>N/A</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>Electrostatics and electricity</td>
<td>3</td>
<td>0.60</td>
<td>5</td>
<td>1.00</td>
</tr>
<tr>
<td>Optics</td>
<td>0</td>
<td>N/A</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>Simple harmonic motion and waves</td>
<td>0</td>
<td>N/A</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>Magnetism</td>
<td>2</td>
<td>0.50</td>
<td>2</td>
<td>1.00\textsuperscript{b}</td>
</tr>
<tr>
<td>Modern physics and electronics</td>
<td>1</td>
<td>1.00</td>
<td>3</td>
<td>1.00\textsuperscript{b}</td>
</tr>
</tbody>
</table>

\textsuperscript{a}GPT: Generative Pre-trained Transformers.
\textsuperscript{b}Highest combination of accurate responses and accuracy consensus in a topic.
\textsuperscript{c}N/A: not applicable.

Scoring Performance

Overall Scores
GPT-4 achieved the highest score with 300 (42.9\%) out of 700 points, outperforming GPT-3.5, which scored 145 (20.7\%) out of 700 points, and Bard, which obtained 115 (16.4\%) out of 700 points. To qualify for the NEET-2023 entrance test, candidates needed to secure at least 137 out of 720 points, which represents 19.6\% of the total points. It was evident that GPT-4 passed the entrance test with flying colors, showcasing exceptional performance. On the other hand, GPT-3.5 managed to meet the qualifying criteria, but with a substantially lower score. However, Bard failed to meet the qualifying criteria and, hence, did not pass the test.

Subject-Level Scores
The subject-level scores, as per NEET’s grading rubric, are detailed in Table 5. GPT-4 achieved the highest overall score of 42.9\% (300/700), outperforming both GPT-3.5 (145/700, 20.7\%) and Bard (115/700, 16.4\%). In all 3 subjects, GPT-4 obtained the highest scores. GPT-3.5 scored higher than Bard in biology and physics but tied with Bard in chemistry.
We then analyzed the breakdown of the total scores obtained by Bard, GPT-3.5, and GPT-4, categorized by subject. Of the total GPT-4 score, 50.3% (151/300) came from biology, 35% (105/300) came from physics, and 14.7% (44/300) came from chemistry. For GPT-3.5, biology contributed 49% (71/145) of the score, physics contributed 34.5% (50/145), and chemistry contributed 16.6% (24/145). Lastly, Bard’s score breakdown showed that 48.7% (56/115) from biology, 30.4% (35/115) came from physics, and 20.9% (24/115) came from chemistry. These results show that GPT-4 outperformed both GPT-3.5 and Bard in the NEET grading rubric, achieving the highest overall score and the top scores in each individual subject. While GPT-3.5 demonstrated better performance than Bard in biology and physics, it tied with Bard in chemistry. The breakdown of scores by subject revealed that for all 3 models, the largest portion of their scores came from biology (understandably, because there were twice as many questions in this category), followed by physics, and then chemistry, indicating a consistent pattern in their relative strengths across these subjects.

**Topic-Level Scores**

The results in Table 5 shows that GPT-4 exhibited strong performance across all topics in physics but showed a relative weakness in inorganic chemistry within the chemistry subject. Bard, compared to the GPT versions, excelled specifically in the biology topics of the plant kingdom and ecosystem and environmental issues.

Table 5. Subject and topic level scores for Bard, Generative Pre-trained Transformers (GPT)-3.5, and GPT-4.

<table>
<thead>
<tr>
<th>Subject and topic</th>
<th>Scores obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bard</td>
</tr>
<tr>
<td>Overall (n=700), n (%)</td>
<td>115 (16.4%)(^a)</td>
</tr>
<tr>
<td>Biology (n=396)</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>56(^a)</td>
</tr>
<tr>
<td>Animal kingdom</td>
<td>6(^a)</td>
</tr>
<tr>
<td>Plant kingdom</td>
<td>30(^b)</td>
</tr>
<tr>
<td>Ecosystem and environmental issues</td>
<td>10(^b)</td>
</tr>
<tr>
<td>Cell biology and genetics</td>
<td>−1(^a)</td>
</tr>
<tr>
<td>Biomolecules</td>
<td>2(^b)</td>
</tr>
<tr>
<td>Population and ecology</td>
<td>−1(^b)</td>
</tr>
<tr>
<td>Evolution and health</td>
<td>1(^a)</td>
</tr>
<tr>
<td>Biotechnology</td>
<td>9(^a)</td>
</tr>
<tr>
<td>Chemistry (n=160)</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>24</td>
</tr>
<tr>
<td>Inorganic chemistry</td>
<td>15(^a)</td>
</tr>
<tr>
<td>Organic chemistry</td>
<td>1</td>
</tr>
<tr>
<td>Physical chemistry</td>
<td>8</td>
</tr>
<tr>
<td>Physics (n=144)</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>35(^a)</td>
</tr>
<tr>
<td>Modern physics and electronics</td>
<td>−4(^a)</td>
</tr>
<tr>
<td>Magnetism</td>
<td>−1(^a)</td>
</tr>
<tr>
<td>Simple harmonic motion and waves</td>
<td>4(^b)</td>
</tr>
<tr>
<td>Optics</td>
<td>−3(^a)</td>
</tr>
<tr>
<td>Electrostatics and electricity</td>
<td>19</td>
</tr>
<tr>
<td>Heat and thermodynamics</td>
<td>2(^b)</td>
</tr>
<tr>
<td>Mechanics</td>
<td>18</td>
</tr>
</tbody>
</table>

\(^a\)Lowest scorer within the topic.
\(^b\)Top scorer within the topic.
environmental issues. Both GPT models performed equally well in cell biology and genetics (biology) and in modern physics and electronics (physics). Additionally, GPT-3.5 stood out for its excellent performance in inorganic chemistry, highlighting its strength in this area of the chemistry subject.

**Discussion**

**Overview**

We evaluated the decision-making performance of 3 models—Bard, GPT, and GPT-4—using accuracy, accuracy consensus, and test scores for the NEET-2023 entrance test. Subject-wise and topic-wise analyses were also conducted. GPT-4 consistently outperformed Bard and GPT across all subjects, achieving the highest accuracy rates: 73% (29/40) in physics, 44% (16/36) in chemistry, and 51% (50/99) in biology. Topic-wise comparisons also demonstrated GPT-4’s excellence in 15 (79%) out of 19 topics, with Bard and GPT excelling in certain topics. Particularly, Bard excelled in simple harmonic motion and waves, while GPT showed strength in inorganic chemistry. Overall, GPT-4 emerged as the top performer, excelling in both subjects and specific topics. Our findings are in line with previous studies that have also examined how LLMs perform on exams related to medical education. Bommineni et al [29] found that GPT-3.5 performs at or above the median performance of the Medical College Admissions Test takers. Ali et al [33] reported that GPT-4 outperformed both GPT-3.5 and Bard by achieving the highest score of 82.6% in specialized questions prepared for neurosurgery oral board examinations. Friederichs et al [34] found that GPT-3.5 answered about two-thirds of the multiple-choice questions correctly and outperformed nearly all medical students in years 1-3 of their studies. Gilson et al [30] reported that GPT-3.5’s performance on the USMLE was either at or near the minimum passing threshold, even without domain-specific fine-tuning. Below, we present both practical and research implications of our findings to enrich the existing literature.

**Implications**

**Practical Implications**

The findings have important implications for users who need to select a model based on specific requirements and their desired score. The subject- and topic-level scores highlight the suitability of different models for various subjects. GPT-4 appears to have the highest score (300/700, 42.9%), followed by GPT-3.5 (145/700, 20.7%), and then Bard (115/700, 16.4%). This demonstrates that Bard was not able to pass the NEET-2023 admission exam, and GPT-3.5 was only 2% (14/700) away from the cutoff score, which is 19% (133/700).

Although GPT-4 appears to be the preferred choice for NEET preparation, it is important to note that GPT-4 is a subscription-based service and the pricing model is uniform across the globe, which makes this model less accessible to the general audience in some parts of the world, particularly low-income countries. When cost is an issue, prospective medical school students might consider using GPT-3.5 and Bard in tandem to develop specialized knowledge and expertise in specific subject topics. The accuracy consensus metric demonstrates that the duo was correct on 29 (42%) out of 69 matching responses, reaching 16.6% (29/175) overall accuracy. However, this duo did not excel in any of the subjects, compared to the other duos. Moreover, at the topic level, it only excelled in “evolution and health.” These results suggest that, in the absence of GPT-4, while students may consider both GPT-3.5 and Bard together for exam preparation, due to the low level of consensus between these models, the total score would still fall below the cutoff score. Moreover, students would be more often confused about the correct responses while cross-checking answers with these models. Therefore, it is recommended that, for exam preparation, students do not solely rely on these models or model duos; instead, they should consult primary sources in conjunction with these models.

**Research Implications**

While scoring performance comparisons help us evaluate whether these models are able to ace the NEET-2023 exam or not, prediction performance comparisons help us evaluate their long-term performance beyond NEET 2023. The models’ predictive accuracy rates match their scoring performance. GPT-4 demonstrated the highest accuracy rate among the 3 models, indicating its superior capability to provide correct responses and its reliability as an accurate study partner. However, there is still plenty of room for improvement since its accuracy was only at 54.3% (95/175), suggesting that anyone using this model for exam preparation would be exposed to a little over 50% (100/200) of accurate information. GPT-3.5 (64/175, 37.6%) and Bard (58/175, 33.1%) had similar overall accuracy rates that are much lower compared to GPT-4’s, suggesting that these 2 models would require significant fine-tuning to qualify as reliable study aids for NEET.

The subject- and topic-level accuracy comparisons highlight specific areas where these models could benefit from domain-specific enhancements. GPT-4 demonstrated superior accuracy across all 3 subjects and 15 topics but required further improvements in 3 topics, that is, ecosystem and environmental issues, plant kingdom, and inorganic chemistry. GPT-4 excelled in at least 1 topic from each subject category, including simple harmonic motion and waves and optics in physics, physical chemistry in chemistry, and evolution and biotechnology in biology. Bard excelled in simple harmonic motion and waves, and GPT-3.5 notably excelled in inorganic chemistry. GPT-3.5, besides requiring improvements in its overall prediction capabilities, needs to develop predictive expertise in population and ecology (biology) and simple harmonic motion and waves (physics). Similarly, Bard needs to develop predictive capabilities in modern physics and electronics and optics, in addition to requiring substantial enhancements in its overall predictive capabilities.

In summary, the implications and applications of this study on LLM and education are far-reaching. First, it could serve as a benchmark for evaluating and improving LLMs’ performance in exams and other educational tasks, enhancing the overall effectiveness of these models in educational settings. Second, the use of LLMs as tutors, mentors, or peers has the potential to significantly enhance students’ learning outcomes and motivation, particularly in a country such as India with a vast
student population and diverse learning needs. Last, this approach could serve as a platform to explore and address ethical and social concerns related to LLMs in education, such as issues of fairness, bias, privacy, and accountability, ensuring responsible and informed use of these models in educational contexts.

Limitations and Further Research
Similar to any other research, this study has certain limitations that should be considered carefully. It is important to note that this study did not involve direct input from actual students, teachers, or medical school boards to understand their perspectives on these mainstream LLMs’ capability to answer questions on basic science concepts. Moreover, we do not know how prospective examinees are using these models for exam preparation or whether they trust them for critical issues such as exam preparation.

LLMs have evolved considerably just in the last 6 months. Therefore, the results of this study will have to be revisited at a later stage. For example, it is possible (and likely) that the relative performance of the different models will change over time. While Bard is currently lagging GPT-3.5 in this area, improvements to the model could mean that it might catch up to GPT-3.5 in the future. Since there is currently an “AI race” among many technology firms, it is only a matter of time before new models are introduced that could perform better on these types of questions.

Conclusion
In this study, we conducted a comparative analysis of 3 notable chatbots, Bard, GPT-3.5, and GPT-4, to evaluate their performance on NEET-2023, a highly competitive medical school entrance examination in India. The study involved the preparation of NEET-2023 questions for the chatbots, data collection, data analysis, and scoring performance assessments.

Our results indicate that GPT-4 not only passed the NEET-2023 entrance test with a score of 42.9% (300/700) but also demonstrated higher accuracy and consensus compared to both GPT-3.5 and Bard. Particularly, GPT-4 consistently outperformed the other models across subjects and topics, achieving an overall accuracy of approximately 54.3% (95/175). GPT-3.5 and Bard, on the other hand, showed variations in their performances, with specific strengths in certain subjects and topics. Regarding subject-wise scoring, GPT-4 excelled in physics and biology while Bard performed well in chemistry. These findings shed light on the proficiency of LLMs in answering high-stakes examination questions, particularly in the context of medical entrance exams such as the NEET. GPT-4’s superior performance and accuracy suggest its potential utility as a valuable resource for medical students seeking assistance in test preparation and knowledge acquisition. However, it is essential to note that despite their impressive performance, LLMs such as Bard, GPT-3.5 and GPT-4 can sometimes generate text containing fabricated or incorrect information. This raises concerns about the credibility of information produced by LLMs, especially in educational settings where accuracy is crucial.

It is also important to acknowledge that LLMs, including GPT, come with both positive and negative consequences [37,38]. Friederichs et al [34] argue that the ability to acquire knowledge is a basic determinant of a physician’s performance, and GPT-3.5 should be looked upon as a tool that provides easy access to a lot of relevant information, eventually aiding in clinical decision-making processes. On the other hand, Mbakwe et al [39] have commented that GPT-3.5’s success on exams such as the USMLE demonstrates the flaws of medical education, which is “mostly focused on the rote memorization of mechanistic models of health and disease” and does not reward critical thinking to the same extent.

Further research and development are warranted to address the limitations and challenges posed by LLMs and ensure their reliable and accurate use in education and other domains. Moreover, future investigations can explore the suitability of LLMs for addressing the needs of diverse professional fields beyond medical entrance exams.

In conclusion, this study contributes valuable insights into the capabilities of Bard, GPT-3.5, and GPT-4 in handling high-stakes examination questions. As LLMs continue to evolve, their potential to revolutionize education and other industries remains promising, albeit with the need for continuous improvements and validation of their accuracy and reliability.

Data Availability
Data can be obtained through a reasonable request to the corresponding author.

Authors’ Contributions
FF and SSS contributed to conceptualization. FF and DØM performed the data acquisition. FF and BMC performed the data analysis. FF, BMC, MN, and SSS contributed to writing and drafting. BMC, MN and DØM contributed to reviewing and proofreading. SSS was the collaborative lead.

Conflicts of Interest
None declared.

References

https://mededu.jmir.org/2024/1/e51523


Abbreviations

AI: artificial intelligence
FP: false positive
GPT: Generative Pre-trained Transformers
LLM: large language model
NEET: National Eligibility cum Entrance Test
NTA: National Testing Agency
TP: true positive
USMLE: United States Medical Licensing Examination

Edited by T Leung, T de Azevedo Cardoso; submitted 02.08.23; peer-reviewed by R Odabashian, M Májovský; comments to author 05.09.23; revised version received 22.09.23; accepted 30.10.23; published 21.02.24.

Please cite as:
Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen DO
Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard
JMIR Med Educ 2024;10:e51523
URL: https://mededu.jmir.org/2024/1/e51523
doi:10.2196/51523
PMID:38381486
Correction: How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment

Aidan Gilson, BS; Conrad W Safranek, BS; Thomas Huang, BS; Vimig Socrates, MS; Ling Chi, BSE; Richard Andrew Taylor, MD, MHS; David Chartash, PhD

Section for Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, CT, United States
Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT, United States
Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States
School of Medicine, University College Dublin, National University of Ireland, Dublin, Dublin, Ireland

Corresponding Author:
David Chartash, PhD
Section for Biomedical Informatics and Data Science
Yale University School of Medicine
300 George Street
Suite 501
New Haven, CT, 06511
United States
Phone: 1 203 737 5379
Email: david.chartash@yale.edu

Related Article:
Correction of: https://mededu.jmir.org/2023/1/e45312
doi:10.2196/57594

In “How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge” (JMIR Med Educ 2023;9:e45312) three additions were made to enhance discoverability.

The title originally appeared as:
How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment

And has been changed to:
How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge

In the “Objective” section of the Abstract, the following sentence:

This study aimed to evaluate the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination Step 1 and Step 2 exams, as well as to analyze responses for user interpretability.

Has been changed to read as:

This study aimed to evaluate the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams, as well as to analyze responses for user interpretability.

Finally, the abbreviation “USMLE” has been added to the Keywords section.

The correction will appear in the online version of the paper on the JMIR Publications website on February 27, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.
ChatGPT in Medical Education: A Precursor for Automation Bias?

Tina Nguyen¹, PhD
The University of Texas Medical Branch, Galveston, TX, United States

Corresponding Author:
Tina Nguyen, PhD
The University of Texas Medical Branch
301 University Blvd
Galveston, TX, 77551
United States
Phone: 1 4097721118
Email: nguyent921@gmail.com

Abstract
Artificial intelligence (AI) in health care has the promise of providing accurate and efficient results. However, AI can also be a black box, where the logic behind its results is nonrational. There are concerns if these questionable results are used in patient care. As physicians have the duty to provide care based on their clinical judgment in addition to their patients’ values and preferences, it is crucial that physicians validate the results from AI. Yet, there are some physicians who exhibit a phenomenon known as automation bias, where there is an assumption from the user that AI is always right. This is a dangerous mindset, as users exhibiting automation bias will not validate the results, given their trust in AI systems. Several factors impact a user’s susceptibility to automation bias, such as inexperience or being born in the digital age. In this editorial, I argue that these factors and a lack of AI education in the medical school curriculum cause automation bias. I also explore the harms of automation bias and why prospective physicians need to be vigilant when using AI. Furthermore, it is important to consider what attitudes are being taught to students when introducing ChatGPT, which could be some students’ first time using AI, prior to their use of AI in the clinical setting. Therefore, in attempts to avoid the problem of automation bias in the long-term, in addition to incorporating AI education into the curriculum, as is necessary, the use of ChatGPT in medical education should be limited to certain tasks. Otherwise, having no constraints on what ChatGPT should be used for could lead to automation bias.

(JMIR Med Educ 2024;10:e50174) doi:10.2196/50174

KEYWORDS
ChatGPT; artificial intelligence; AI; medical students; residents; medical school curriculum; medical education; automation bias; large language models; LLMs; bias

Introduction
With the introduction of artificial intelligence (AI), automated processes for nearly most tasks have become the norm. In the clinical environment, AI has been used for diagnosis, prognosis, and administrative tasks. Given the popularity of other forms of AI—as seen most recently with ChatGPT, a large language model developed by the company OpenAI—there are suggestions for its potential role in medical education. Users of ChatGPT boast its efficiency and relative accuracy, such as in the generation of a patient’s discharge summary or the conduction of literature reviews [1]. As advancements in medicine continue to arise, medical students are burdened with the impossible task of balancing the need to continuously learn and retain competencies and the need to provide compassionate patient care. As a result, some medical students might feel an incentive to use ChatGPT to save them time in their busy schedules. However, despite the novel acclaim, the technical and ethical issues seen with AI, such as biased results or nonsensical outputs, also plague ChatGPT. These problems become exacerbated when medical students inadvertently develop automation bias, where they overrely on AI, and continue to have this mentality when they become residents, at which point they have the potential to harm patients if the AI provides an erroneous outcome. In this editorial, I argue the justification for AI education in the medical school curriculum and how the lack of it leads to the problem of automation bias, as well as the other harms from automation bias. Subsequently, I connect the implications of students using ChatGPT with automation bias. Finally, I provide recommendations for when ChatGPT use is appropriate.
The Need for AI Education in the Medical School Curriculum

As the health care landscape has drastically changed through the years, physicians have had to quickly adapt to the digital age. Given the amount of information physicians are required to retain and the new information they must continue to learn, such as information on emerging diseases and the health data of the patients they track, physicians are expected to interact with computer systems in some capacity, whether it is for charting their patients’ information or consulting clinical decision support systems. However, the lack of content on the technological systems in the health care setting inhibits prospective physicians from understanding the benefits of using these technologies, the ethical issues that can arise with their use, and future innovations, along with the wider implications of AI. In Civaner et al’s [2] survey of medical students’ opinions on AI education, they found that 75.6% of students had either limited or no education on the topic of AI. These participants also noted not feeling well equipped to work with AI in the clinical setting. Additionally, in Yun et al’s [3] proposal for future internal medicine physicians, they suggested that these prospective physicians should be able to appreciate the roles of big data and AI in health care. Clearly, there is a desire from students, as well as residency and fellowship programs, to incorporate AI education into the medical school curriculum and training. AI education and training cannot continue to be delayed, as some forms of AI have already been deployed in the clinical setting.

Although several studies have provided proposals for implementing AI education into the medical school curriculum, they have also noted the difficulties of developing AI education, such as schedule constraints and the challenges of deciding the material that should be covered [4,5]. Additionally, this task should not solely be deferred to the attending physicians, as they themselves might not have the adequate training with AI to teach others [5]. Although these challenges serve as barriers to implementing quality AI education into the curriculum, an attempt to include at least some type of education on or educational resources about AI is needed to prepare students and potentially prevent problems in the clinical setting, as further explored in the following section. Therefore, future physicians, medical students, and residents should be trained on the use of AI in health care and other related topics, such as big data or machine learning, to understand the tools they will be working with. Even though medical students should not be expected to be experts in AI and know every technical aspect of these technologies, they should at least feel comfortable with navigating how and when to use AI.

The Problem of Automation Bias

Although AI is supposed to aid physicians in various processes to decrease their workload and give them more time with their patients, AI can also cause unintended ethical issues. One of the common ethical concerns with AI is that it can essentially be a black box, where the results from the AI are illogical, and the AI developer cannot track how it produced those erroneous results. This problem becomes exacerbated when automation bias arises. Automation bias occurs when a user overrelies on AI systems. Therefore, if a physician exhibits automation bias, then they will not question the results from the AI, potentially leading to bad medical care. In Lyell et al’s [6] study, the error rate associated with a clinical decision support system when it was inaccurate was higher (86.6%) in comparison to the rate it had when it was accurate (58.8%). Although automated processes aid in decision-making and can provide accurate results, there is also the possibility of these systems providing incorrect results and causing irreversible harm on a much larger scale. An example includes the Prescription Drug Monitoring Program (PDMP), a machine learning system that provides risk scores for patients’ likelihood to misuse prescription drugs, which can cause both testimonial injustice and physical harm [7,8]. Testimonial injustice, a form of epistemic injustice, develops when a patient’s account of their health is unfairly dismissed by their physician [8]. Testimonial injustice invalidates the credibility of patients and further implies that their care is dependent on how physicians deem their trustworthiness [8]. A patient’s risk scores can be negatively affected if their chart becomes commingled, which is also known as overlay, where a specific person’s electronic health record erroneously pulls in the data of other patients with similar demographic characteristics and compiles these data into 1 chart [7,9]. As such, a patient with chronic pain may not receive the medication they need due to the PDMP providing an incorrect risk score. If a physician uses the risk scores of the PDMP without validating the results or considering their patients’ testimonies, then physical harm, as well as patients’ mistrust toward the physician and the potential deterrence of seeking health care, will ensue. Although AI can aid in the decision-making process, ultimately it is the duty of the physician to ensure that their decisions are based on sound clinical judgment. As such, if a physician with automation bias applies an erroneous outcome to a patient’s case, then the physician becomes accountable for that outcome instead of the AI, as they are the party that used the outcome. To clarify, more sophisticated AI and machine learning systems have been proposed, of which the results would be difficult for users to verify, as these systems use advanced techniques that do not rely on predefined rules. However, the AI systems described in this section are known as expert systems, which use a coded set of rules and rely on predefined rules [10]. Even though the verification process might essentially be beyond the scope of some physicians’ expertise regarding future AI and machine learning, physicians should remain attentive to results from AI.

The Implications for Medical Students and Residents

As seen with the case of the PDMP, automation bias can lead to various harms. Therefore, the systemic issue of automation bias in health care must be addressed. The mentality that AI is always right is often associated with medical students and residents [6,11]. As these groups have grown up in the digital age, they are more comfortable with embracing technology into their practice than older physicians (who either lack digital literacy or are resistant to change). In addition to their openness
to using AI, medical students and residents might be prone to automation bias, as they lack experience or are not confident in their skills [11]. Multiple studies have found that algorithmic appreciation—a user’s valuing of an algorithm’s outputs—is lower for users who have more experience in a task than for those who are considered nonexperts in that task [12,13]. A combination of factors, such as newer physicians being digital natives, insufficient expertise, and less overall confidence, highlights how the systemic problem of automation bias came to be. Therefore, the deficiency of AI education in medical school and beyond sets up users to become susceptible to automation bias, as they might be unaware of the technical problems with AI. These users will come into the clinical setting with the assumption that AI systems are always accurate, which will cloud their clinical judgment.

In addition to the broader discussion of AI in health care, which students will inevitably have to interact with at some point in their professional careers, I want to focus on an AI that is accessible to students now—ChatGPT. The fact that ChatGPT has passed the US Medical Licensing Examination could entice students to use ChatGPT [14]. Moreover, Tiwari et al [15], who applied the Technology Acceptance Model to ChatGPT, found that students generally had positive views (in terms of perceived usefulness, credibility, social presence, and hedonic motivation) of ChatGPT based on their previous experiences with using the tool. However, just as AI can be a black-box algorithm, so too can ChatGPT, with respect to its hallucinations. ChatGPT’s hallucinations are results that are seemingly feasible but do not actually exist [1,16]. For example, it is commonly known that ChatGPT can make up citations [16,17]. Additionally, in an editorial, ChatGPT had to be prompted several times by the author to finally respond that it cannot generate visual diagrams [18]. Further, ChatGPT’s data sources only cover data from 2021 and prior years, and as its scope is limited to this context, ChatGPT can provide outdated information [19]. Therefore, despite the acclaim, ChatGPT is not as perfect as some claim it to be. Given the push for ChatGPT use, there is a risk that users might develop an AI solutionism mentality, where users assume that AI has the answer to all problems [10]. AI solutionism is closely related to automation bias, as users with the preconceived notion that AI is always right are more willing to turn to AI. As such, if we train medical students to use ChatGPT, will they be more predisposed to automation bias in the future when they become residents? Although there is no direct answer to this question, given what is known about the medical school curriculum, the context of the student population being composed of digital natives, and the AI solutionism mentality, the possibility of this happening seems likely. Some medical students will take their past, positive interactions with ChatGPT, wherein they received the right response, as confirmation that ChatGPT is reliable. The concern here is that students’ perceptions of the reliability of ChatGPT dictate their views on AI, including AI in the clinical setting, making it easier for them to become susceptible to automation bias. Although some suggest using AI suppression, an approach where an AI’s recommendations are not provided if there is “a higher misleading probability,” to mitigate the risk of automation bias, there appears to be no concrete solutions to solving this problem, especially in the context of the “novice” medical student and resident population [20]. It must also be acknowledged that sometimes, AI use cannot be completely avoided in the health care setting. Thus, in controlling the reoccurrence of automation bias, I believe that students must not only be aware of this potential problem but also build the skills required to prevent this mentality. When addressing the risks of AI in the medical school curriculum, automation bias needs to be a discussion topic. Besides teaching about automation bias, when training medical students, it is important to consider the “hidden curriculum” about using AI, that is, the implied lessons, cultures, and views that students learn in lectures or from observations of faculty [21]. If faculty also fall into the trap of AI solutionism, this will lead to a biased perspective on AI and contribute to the “hidden curriculum.” Faculty should serve as an example for students by ensuring that students have the right critical analysis skills and are comfortable with questioning results instead of accepting what is being given to them. This builds students’ confidence in trusting their instincts, which could deter them from automation bias.

When Should ChatGPT Be Used in Medical Schools?

Although this editorial takes a more critical stance on AI and ChatGPT, I want to clarify that this does not mean that these tools should never be used or that their functionalities are ineffective. Notably, in the preclinical phase, the medical school curriculum is not catered to students, as the focus is on ensuring that students have expertise on basic medical concepts, the structure and functions of the body, diseases, diagnoses, and treatment concepts [22,23]. This might be a challenge for some students who prefer different learning methods as opposed to the typical didactic method. ChatGPT can be a beneficial tool for students who prefer student-centered or self-directed learning, as it excels in summarizing information and generating practice questions [18,19,24,25]. Students who struggle with a concept in class or want further explanations could also use ChatGPT as an additional resource. Being able to personalize their learning experiences encourages students toward incorporating ChatGPT into their studies. As such, banning the use of ChatGPT could result in students being even more enticed to seek out the “forbidden” chatbot. Therefore, in addition to integrating AI education into the medical school curriculum and avoiding the “hidden curriculum” about AI, students should feel encouraged to use ChatGPT but only to a certain extent. Despite the advantages of ChatGPT use, students should not be compelled to turn to ChatGPT for every task. For example, assignments that involve students writing about their firsthand experiences would not be appropriate for ChatGPT. With regard to a hypothetical student who delegated such an assignment to ChatGPT, van de Riddèr et al [26] stated that “[r]eflections contribute to a learner’s professional development, but this learner robbed themself of an innate self-reflective opportunity.” Students lose a potential outlet for their emotions and the humanistic aspect of care when they delegate ChatGPT to the task of writing a self-reflection piece [27]. Notably, ChatGPT appears to be popular in the context of scientific writing for the following reasons: “efficiency and versatility in writing with
text of high quality, improved language, readability, and translation promoting research equity, and accelerated literature review” [1]. However, Blanco-Gonzalez et al [28] argue that “…ChatGPT is not a useful tool for writing reliable scientific texts without strong human intervention. It lacks the knowledge and expertise necessary to accurately and adequately convey complex scientific concepts and information.” There are also concerns about plagiarism with ChatGPT, as it can fabricate citations, fail to disclose all references, and provide inaccurate content (as it only uses information from 2021 and prior years) [1,17]. Therefore, ChatGPT should not be used for writing, as it deprives students of the opportunity to engage in their professional identity and, for those wanting to go into research, the necessary research skills to conduct empirical or conceptual work. Additionally, some web-based educational resources, such as modules or augmented reality, might help supplement students’ experiences during the clinical phase [29]. However, the use of these resources, including ChatGPT, should not be the only learning experience that students have in the clinical phase. In order to build their interpersonal skills and practice humanistic care, students must interact with real patients and other professionals in the clinical setting. Although some students might feel prepared for these interactions (based on their experiences of working through case scenarios that ChatGPT generated for them), they will soon realize that they cannot predict or account for how patients or others (eg, a patient’s family, members of the care team, etc) react in real time. Learning to accommodate patients’ needs and working in a team cannot realistically be achieved with ChatGPT. Instead, these skills are cultivated through students’ experiences in the clinical setting. The focus should not be on deciding whether to use ChatGPT but on determining the best contexts that ChatGPT can be applied to. As seen in this editorial, ChatGPT excels at particular tasks, such as summarizing information and creating study materials [18,19,24]. Ideally, students should use ChatGPT to supplement their learning experience rather than use it as their sole resource for medical science education. Students should still validate the results (to the extent that they can) from ChatGPT, because it can provide inaccurate results and the problem of hallucinations persists, before they wholeheartedly study or apply the wrong information. When used in this context, ChatGPT plays a lesser role in students’ education, thereby further enhancing their ability to discern results and avoiding AI solutionism.

**Conclusion**

To minimize the risk of students developing automation bias, we need to ensure that students receive proper AI education, in which the courses and lessons will teach them about the ethical issues surrounding AI technologies, as well as the problem of automation bias, and encourage the moderate use of AI. ChatGPT should only be used for certain tasks, and it should not be the default resource that students turn to, as this could cause a domino effect, where students develop the automation bias mentality as a result of developing the AI solutionism mentality. Therefore, training medical students to avoid falling into these traps of AI solutionism and automation bias starts in the classroom. Again, the medical school curriculum must reflect the current needs of the students. Furthermore, faculty serve as an example for students; therefore, they should also be proactive in deterring the use of ChatGPT for all tasks and be careful not to contribute to the “hidden curriculum” about AI. Overall, ChatGPT is an assistive tool but only when used in the right context.

**Acknowledgments**

The author declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

**Conflicts of Interest**

None declared.

**References**


Abbreviations

AI: artificial intelligence
PDMP: Prescription Drug Monitoring Program
Impact of the COVID-19 Pandemic on Medical Grand Rounds Attendance: Comparison of In-Person and Remote Conferences

Ken Monahan\(^1\), MD; Edward Gould\(^1\), MD; Todd Rice\(^1\), MD; Patty Wright\(^1\), MD; Eduard Vasilevskis\(^1\), MD; Frank Harrell\(^1\), PhD; Monique Drago\(^1\), EdD; Sarah Mitchell\(^1\), MS

Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Ken Monahan, MD
Vanderbilt University Medical Center
1215 21st Avenue
Medical Center East - 5th Floor
Nashville, TN, 37232
United States
Phone: 1 6153222318
Email: ken.monahan@vumc.org

Abstract

Background: Many academic medical centers transitioned from in-person to remote conferences due to the COVID-19 pandemic, but the impact on faculty attendance is unknown.

Objective: This study aims to evaluate changes in attendance at medical grand rounds (MGR) following the transition from an in-person to remote format and as a function of the COVID-19 census at Vanderbilt Medical Center.

Methods: We obtained the faculty attendee characteristics from Department of Medicine records. Attendance was recorded using a SMS text message–based system. The daily COVID-19 census was recorded independently by hospital administration. The main attendance metric was the proportion of eligible faculty that attended each MGR. Comparisons were made for the entire cohort and for individual faculty.

Results: The observation period was from March 2019 to June 2021 and included 101 MGR conferences with more than 600 eligible faculty. Overall attendance was unchanged during the in-person and remote formats (12,536/25,808, 48.6% vs 16,727/32,680, 51.2%; \(P=0.44\)) and did not change significantly during a surge in the COVID-19 census. Individual faculty members attendance rates varied widely. Absolute differences between formats were less than –20% or greater than 20% for one-third (160/476, 33.6%) of faculty. Pulmonary or critical care faculty attendance increased during the remote format compared to in person (1450/2616, 55.4% vs 1004/2045, 49.1%; \(P<0.001\)). A cloud-based digital archive of MGR lectures was accessed by <1% of faculty per conference.

Conclusions: Overall faculty attendance at MGR did not change following the transition to a remote format, regardless of the COVID-19 census, but individual attendance habits fluctuated in a bidirectional manner. Incentivizing the use of a digital archive may represent an opportunity to increase faculty consumption of MGR.

(JMIR Med Educ 2024;10:e43705) doi:10.2196/43705

KEYWORDS

continuing medical education; COVID-19; distance education; professional development; virtual learning

Introduction

Medical grand rounds (MGR) has evolved from the bedside [1] to a weekly presentation to the entire department [2]. Due to the COVID-19 pandemic, the format of MGR has undergone another transition, from in person to remote. While MGR attendance patterns for in-person conferences have been reported [3], the impact of remote conferences on faculty attendance at MGR is unknown. The analysis of remote surgical conferences [4,5] has been limited by sample size and aggregate data.

We propose that including more faculty from multiple specialties and individual conference or attendee data will provide more robust analysis that may inform returning to an in-person format, maintaining a remote format, or using a hybrid approach. Therefore, using our institution’s cloud-based attendance recording database, we (1) evaluated MGR attendance over time...
before and after the transition to the remote format and (2) assessed the temporal relationship between our institution’s COVID-19 census and attendance at MGR conferences.

Methods

Study Design, Participants, and Setting
We performed a retrospective cohort study of MGR attendance for all Department of Medicine (DOM) clinical faculty at Vanderbilt Medical Center active between March 2019 and June 2021. All conferences before March 12, 2020, were in person, and all conferences on or following this date were remote.

Attendee Characteristics
For each division within the DOM, the number of faculty eligible to attend each conference as well as the number of faculty that attended each conference were available, as was each faculty member’s academic rank (assistant, associate, or full professor).

Recording of Conference Attendance
Attendance was recorded by a cloud-based continuing medical education (CME) system during the entire observation period. Faculty indicate their attendance by sending an SMS text message containing the unique numeric code for that conference to a specific CME number. Conference attendance is registered as a binary outcome. The number of faculty considered to have attended a conference was obtained directly from this system. The number of faculty considered not to have attended was defined as the difference between the number of faculty eligible to attend and the number for whom attendance was recorded. The proportion of attendance was defined as the ratio of those who attended to those who were eligible over a given time frame (ie, in person or remote).

Individual-Level Attendance Data
For each faculty member, the CME system generates a unique user number that is not related to any other identification mechanism or coupled to any other database. By removing all identifying information from faculty members’ attendance data except this user number, we could track individual attendance over time without the capability of linking these data to a given faculty member’s actual identity.

Archived Conferences
Beginning in November 2019, digital recordings became available shortly after each MGR. Attendance credit was not given for consuming MGR in this manner. The number of faculty members that accessed a given MGR and the date on which each faculty member accessed the conference were available from the archive.

Acquisition of COVID-19–Related Data
Our institution tracked the census of hospital inpatients with positive COVID-19 tests as well as the subset of that group that required intensive care unit (ICU) care or mechanical ventilation. The COVID-19 burden on a given day included the total number of COVID-19 patients (cases) relative to the peak observed during the observation period (calculated as cases or peak), the proportion of patients with COVID-19 requiring ICU care relative to the number of cases (calculated as ICU or cases), and the proportion of patients with COVID-19 requiring mechanical ventilation (calculated as ventilator or cases). We defined the “surge” as the interval between December 2020 and January 2021, when COVID-19 cases were at their maximum.

Statistical Analysis
The main analyses compared the attendance rates during the entire in-person and remote periods as well as during the surge. Additional analyses stratified attendance by academic rank. All comparisons were made using the chi-square test in GraphPad Prism (version 9.2.0; GraphPad Software). For individual attendees, the difference between attendance rates at in-person and remote conferences was calculated, as were the characteristics of the resulting distribution.

Ethical Considerations
This investigation was considered nonresearch activity by the Vanderbilt Medical Center’s institutional review board (number 211362). The need for informed consent was waived because of the retrospective nature of the study.

Results

Cohort Characteristics and Overall Attendance Observations
Characteristics of the MGR conferences, speakers, and faculty attendees are displayed in Table 1.
Table 1. Conference and attendee characteristics.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value at the end of the observation (range during observation period)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conferences, n</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total during observation period</td>
<td>101</td>
<td>N/A*</td>
</tr>
<tr>
<td>In person (pre-pandemic)</td>
<td>47</td>
<td>N/A</td>
</tr>
<tr>
<td>Remote (intrapandemic)</td>
<td>54</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Topic, n</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiology</td>
<td>19</td>
<td>N/A</td>
</tr>
<tr>
<td>Endocrine</td>
<td>10</td>
<td>N/A</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>12</td>
<td>N/A</td>
</tr>
<tr>
<td>General internal medicine</td>
<td>15</td>
<td>N/A</td>
</tr>
<tr>
<td>Geriatric medicine</td>
<td>3</td>
<td>N/A</td>
</tr>
<tr>
<td>Hematology or oncology</td>
<td>10</td>
<td>N/A</td>
</tr>
<tr>
<td>Infectious disease</td>
<td>10</td>
<td>N/A</td>
</tr>
<tr>
<td>Nephrology</td>
<td>7</td>
<td>N/A</td>
</tr>
<tr>
<td>Pulmonary or critical care</td>
<td>7</td>
<td>N/A</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Speaker, n</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal</td>
<td>41</td>
<td>N/A</td>
</tr>
<tr>
<td>External</td>
<td>60</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Faculty attendance</strong>, mean (SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total eligible to attend</td>
<td>579 (22)</td>
<td>611 (544-612)</td>
</tr>
<tr>
<td>Cardiology</td>
<td>100 (2)</td>
<td>103 (95-103)</td>
</tr>
<tr>
<td>Endocrine</td>
<td>25 (2)</td>
<td>28 (23-28)</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>41 (2)</td>
<td>43 (38-43)</td>
</tr>
<tr>
<td>General internal medicine</td>
<td>175 (8)</td>
<td>187 (161-187)</td>
</tr>
<tr>
<td>Hematology or oncology</td>
<td>65 (2)</td>
<td>69 (60-69)</td>
</tr>
<tr>
<td>Infectious disease</td>
<td>43 (1)</td>
<td>45 (40-45)</td>
</tr>
<tr>
<td>Nephrology</td>
<td>33 (2)</td>
<td>36 (31-36)</td>
</tr>
<tr>
<td>Pulmonary or critical care</td>
<td>46 (2)</td>
<td>46 (42-49)</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>22 (1)</td>
<td>23 (21-23)</td>
</tr>
<tr>
<td>Assistant professor</td>
<td>328 (16)</td>
<td>349 (279-350)</td>
</tr>
<tr>
<td>Associate professor</td>
<td>107 (1)</td>
<td>109 (105-109)</td>
</tr>
<tr>
<td>Full professor</td>
<td>143 (11)</td>
<td>149 (107-151)</td>
</tr>
</tbody>
</table>

*aN/A: not applicable.

The number of faculty in the subspecialties is fewer than the total due to not listing smaller divisions. Faculty categorized by academic rank may not sum to the total due to a small number of transitions between ranks.

Figure 1A shows (1) the time series of MGR attendance over the entire observation period and the number of times a given MGR was accessed from the cloud-based archive within 1 month of the conference, (2) the concurrent time series of COVID-19 cases as a proportion of the peak number recorded during the observation period, and (3) the time series of COVID-19 cases requiring ICU care and ICU cases requiring mechanical ventilation, both as proportions of the number of COVID-19 cases. Despite increases in remote attendance during the beginning of the pandemic (Figure 1B) and a brief increase as the surge began to subside (Figure 1C), there was no difference in attendance at MGR during the in-person format and the remote format over the entire observation period (12,536/25,808, 48.6% vs 16,727/32,680, 51.2%; *P*=.44). The proportion of faculty accessing the MGR digital archive remained low throughout the observation period, never exceeding 5% for any lecture and often not exceeding 1% (mean 0.7%, SD 1.3%).

https://mededu.jmir.org/2024/1/e43705
Figure 1. Time series of medical grand rounds (MGR) attendance and concurrent COVID-19 burden. (A) The entire observation period, (B) focus on the beginning of the remote format, and (C) focus on the surge. At the onset of the remote format, there is a nonsustained increase in attendance. As the COVID-19 census increased rapidly leading up to the peak census, there was no change in attendance. During the peak of the surge, there is a very small transient reduction in attendance followed by an extremely brief increase in attendance during a period of rapid decline in the COVID-19 census. Access to archived MGR lectures remained low during the entire observation period. ICU: intensive care unit.

MGR attendance stratified by academic rank across the in-person and remote formats is shown in Figure 2. Associate (3249/5788, 56.1% vs 2515/4989, 50.4%; $P<.001$) and full professor (3309/7718, 42.9% vs 2433/6757, 36%; $P<.001$) attendance was higher at MGR during the remote format relative to the in-person format.

Figure 2. Attendance at medical grand rounds stratified by academic rank. Assistant professor attendance was the same regardless of conference format, whereas associate and full professor attendance increased during the remote format relative to in person. *$P<.001$.

Subinterval and Subgroup Analyses

There was no difference in attendance during the surge compared to the 2 months before (October to November 2020; 2071/4218, 49.1% vs 2194/4229, 51.9%; $P=.38$) or 1 year before (December 2019 to January 2020; 2028/3990, 50.8% vs 2194/4229, 51.9%; $P=.34$).

The attendance trends of DOM subspecialties that were particularly impacted by the pandemic are superimposed on the overall DOM trend in Figure 3 for pulmonary or critical care (CC), infectious diseases (ID), and general internal medicine (GIM).
Figure 3. Selected subspecialty attendance trends. There are distinct qualitative patterns of medical grand rounds (MGR) attendance relative to the entire Department of Medicine (DOM) cohort for faculty in (A) pulmonary or critical care (CC), (B) infectious diseases (ID), and (C) general internal medicine (GIM).

Pulmonary or CC attendance during the remote format was higher than during the in-person format (1450/2616, 55.4% vs 1004/2045, 49.1%; \( P < .001 \)). This attendance pattern persisted while cases were rising and peaking during the surge, when demands on this portion of the faculty were likely greater than prepandemic. ID faculty had higher attendance throughout the entire observation period relative to the whole DOM cohort. The GIM faculty consistently attended MGR less frequently than the rest of the DOM cohort, including a sizable decrease during the peak of the surge.

**Individual-Level Analyses**

Data were available for 476 faculty eligible to attend all the MGR during the observation period. As shown in Figure 4A, attendance rates during in-person conferences did not predict attendance rates for remote conferences. As displayed in Figure 4B, the distribution of the absolute difference between remote and in-person attendance rates is relatively symmetric around the null, but outliers at both tails are noted. Attendance decreased by at least 20% for nearly 15% (64/476; 13.4%) of faculty and increased by at least that amount for 20.2% (96/476) of faculty. The distribution of the differences in individual faculty attendance between remote and in-person conferences is shown in Figure 4C, stratified by in-person attendance rates. The distributions of the 2 lowest categories of in-person attendance exhibit positive skewness, while the remaining categories demonstrate negative skewness, indicating that the direction of the changes in individual attendance patterns observed with the transition in conference format varied based on in-person attendance. Lastly, 4.8% (23/476) of faculty exhibited absolute differences of 50% in attendance between formats.
Figure 4. Individual-level attendance at in-person and remote medical grand rounds (MGR). (A) For a given level of attendance at in-person MGR, individual faculty member attendance at remote MGR fluctuated widely. (B) The distribution of the difference in attendance rates between conference formats for the entire cohort is relatively symmetric around the null, as expected given the overall lack of change. Nonetheless, extreme values of changes in attendance at the tails are noted. (C) Faculty that attended in-person MGR less frequently generally increased their attendance at remote MGR, while the opposite response was observed for those that frequently attended in-person MGR. Red bars indicate the mean.

Discussion

Principal Findings

Overall faculty attendance at MGR remained constant regardless of conference format, suggesting no disadvantage to the remote format. In addition, there may be substantial cost savings [6] and beneficial environmental impacts [7] associated with the remote format as it pertains to external speakers, who comprised the majority (60/101, 59.4%) of this cohort.

The increase in attendance of associate and full professors during the remote format may indicate fewer concurrent clinical obligations for these groups compared to their more junior colleagues. COVID-19–related MGR lectures at the beginning of the remote period may have led to the concurrent initial increase in attendance [8], but attendance quickly regressed to the mean, which was maintained even during a subsequent period of rapid rise and peak in COVID-19 burden.

Paradoxically, pulmonary or CC faculty attendance increased during the pandemic. It is possible that the attendance of the subgroup of non-ICU providers within pulmonary or CC may have increased during the pandemic while the attendance of their ICU-based colleagues declined. We speculate that the decreased attendance of the division of GIM was contributed to by lower attendance within the section of hospital medicine, perhaps because of burnout [9].

Individual faculty attendance habits did not remain static in response to the change in conference format. The pandemic or the remote format may have motivated faculty to attend MGR who did not regularly do so, thus taking the place of faculty that were unable to attend due to increased clinical or administrative responsibilities. The presence of outliers at both extremes of attendance shifts may enrich further investigations of specific drivers of conference attendance, which could inform decisions regarding conference format moving forward.

Archived conferences were infrequently accessed throughout the observation period. Encouraging asynchronous viewing may increase consumption of MGR among faculty who are unable to do so in real time. Offering attendance credit for viewing MGR asynchronously could incentivize otherwise nonattending faculty.

Limitations

This study did not use surveys or other methods of obtaining feedback from faculty regarding their attendance patterns relative to the mode of MGR presentation, as collecting these data was not feasible given the study’s retrospective design.

Attendance does not guarantee the observer has learned from MGR, although mandatory evaluations may not assess this objective either [10].

Conclusions

Overall faculty attendance at MGR was neither durably affected by a pandemic-related transition from in-person to a remote format nor by concurrent COVID-19 burden, although individual attendance behaviors varied considerably. If coupled with archived conference recordings, the remote format may be an equally attended and more cost-effective option for presenting MGR in a postpandemic era.
Acknowledgments
The authors wish to thank Attallah Stout and Joseph Braeuner for assistance with medical grand rounds topic, speaker, and attendance data; Ariel Dunham for assistance with the medical grand rounds digital archive; and Brandi Cherry and Chad Fitzgerald for assistance with COVID-19 census data.

Conflicts of Interest
None declared.

References

Abbreviations
CC: critical care
CME: continuing medical education
DOM: Department of Medicine
GIM: general internal medicine
ICU: intensive care unit
ID: infectious diseases
MGR: medical grand rounds
Using AI Text-to-Image Generation to Create Novel Illustrations for Medical Education: Current Limitations as Illustrated by Hypothyroidism and Horner Syndrome

Ajay Kumar¹, MBBS, MSc; Pierce Burr¹, BSc, MBChB, MRCSEd; Tim Michael Young¹, BSc, MBBS, PGCME, PhD

Corresponding Author:
Tim Michael Young, BSc, MBBS, PGCME, PhD
Queen Square Institute of Neurology
University College London
Number 7 Queen Square
London, WC1N 3BG
United Kingdom
Phone: 44 2031082781
Email: t.young@ucl.ac.uk

Abstract

Our research letter investigates the potential, as well as the current limitations, of widely available text-to-image tools in generating images for medical education. We focused on illustrations of important physical signs in the face (for which confidentiality issues in conventional patient photograph use may be a particular concern) that medics should know about, and we used facial images of hypothyroidism and Horner syndrome as examples.

(Methods)

ChatGPT was used to generate prompts for the two AI text-to-image tools used in this study—DALL·E 2 and Midjourney (Multimedia Appendix 1) [8-10], with which the prompts were used to generate images for hypothyroidism and Horner syndrome. The images were assessed and selected, using the following suitability criteria:

1. Images were excluded if any of the following features were present: insufficient coverage of the face, blurred images, a lack of realistic or humanoid features, a lack of continuity of edges, background noise, cloning errors, and geometrical and shadow inconsistencies.

2. Remaining images were accepted if they adequately represented the facial features of hypothyroidism or Horner syndrome, as judged by the coauthors (all were experienced physicians).

If adequate images could not be generated via the above methods, additional prompts, which were not generated with ChatGPT, were used. If adequate images were still not generated, then secondary editing via Microsoft Paint and GNU
Image Manipulation Program (GIMP) was performed on the best image to try and meet the criteria listed above.

Results

Facial Features of Hypothyroidism

Using ChatGPT, the following text prompt was generated (restricted to the DALL·E 2 prompt word limit):

Generate an image depicting a middle-aged Caucasian woman with hypothyroidism presenting with facial myxedema. The woman should be shown in a frontal view, focusing on her face, scalp, and neck, without any makeup. The face must be very rounded and extreme scalp balding with coarse hair. Skin looks dry and pale. Outer eyebrows have a paucity of hairs, eyelids look very puffy. She looks tired.

The prompt was used to generate 120 images. Of these, 53 were removed, using our preset exclusion criteria. Of the remaining 67, only 17 met some of the criteria for adequately representing facial features of hypothyroidism. The best image was selected as Figure 1 [9], with no additional editing needed.

Figure 1. Artificial intelligence text-to-image production of facial features typical of hypothyroidism (myxedema) showing classical clinical features, including a rounded face with dry, pale skin; puffy eyelids; a general appearance of tiredness; and partial balding with coarse hair and loss of hair in the eyebrows (especially in the outer third). This image was produced by using DALL·E 2 [9] alone and without additional editing.

Horner Syndrome

The following prompt was obtained from ChatGPT:

Create an illustrative depiction of a patient displaying Horner’s syndrome, emphasizing the key clinical features, such as ptosis (drooping of the upper eyelid), miosis (constricted pupil), and anhidrosis (lack of sweating) on one side of the face. Ensure the image is clear and medically accurate, aiding in the understanding of this neurological condition.

Of the 120 images, 85 met our exclusion criteria, but none met our inclusion criteria, even after alternative prompts and DALL·E 2 were used. We therefore selected the best image (produced by Midjourney) and then performed secondary editing with Microsoft Paint and GIMP (Figure 2 [10]). This produced an image of Horner syndrome that was judged as adequate.

Figure 2. Generated illustration of Horner syndrome. Image 1 was produced by using Midjourney [10]. Image 2 shows the result after minor image editing (as described in our Methods section) to attenuate the key teaching features, which are labeled in image 3 (A: ptosis; B: miosis; C: apparent enophthalmos; D: upside-down ptosis).
**Discussion**

We aimed to explore the potential, as well as the current limits, of AI text-to-image generation in producing illustrations of medical conditions affecting the face. Without the use of high-quality medical images, it can be more challenging to teach others about these important conditions [11]. We showed that AI text-to-image generation is readily possible for hypothyroidism—a condition with symmetrical features. However, for Horner syndrome—a condition with asymmetrical features—adequate images could only be produced after some additional slight editing, reflecting a possible limiting factor of these tools. Ours are the first AI-generated images of classical facial features of hypothyroidism and Horner syndrome that we are aware of.

Confidentiality has become an increasing concern in the use of medical images over the last few decades. Text-to-image tools have ethical issues, including issues of consent for the original photos used to train these tools. Additionally, issues of accuracy are key. Nonmedics might be misled on medical signs by using such tools. Targets for future research are the potential for biases with these tools and the danger of stereotypes being perpetuated. Despite these limitations, AI-generated images may enhance case-based learning, allowing students to study and analyze a diverse range of medical cases. Text-to-image tools show exciting potential and may allow easier access to high-quality images in medical education [12,13].

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

Tools used in this article (all prompts entered in English).

[DOCX File, 13 KB - mededu_v10i1e52155_app1.docx ]

**References**


**Abbreviations**

AI: artificial intelligence

GIMP: GNU Image Manipulation Program

https://mededu.jmir.org/2024/1/e52155