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, Ruiy Huang, Guan-chun Lin, Jhen-Yu Lin, Clydie Aldana. .................................................. 3

A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study (e53961)
Friederike Holderried, Christian Stegemann–Philips, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling. .................................................. 12

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 Alfterthofer, Samuel Knoedler, Cosima Hoch, Paul Funk, Sebastian Cotofana, Bhagvat Maheta, Konstantin Frank, Vanessa Brébant, Lukas Prantl, Philipp Lamby. .................................................. 43

Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects (e51247)
Lukas Weidener, Michael Fischer. .................................................................................................................. 53

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. .......................................................................................... 71

The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators' Perspectives (e47339)
Yaser Al-Worafi, Kiang Goh, Andi Hermansyah, Ching Tan, Long Ming. .................................................. 86

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. .................................................. 93

Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study (e50842)
Firas Haddad, Joanna Saade. .................................................................................................................. 116
**Viewpoints**

Patients, Doctors, and Chatbots *(e50869)*
Thomas Erren. .................................................................................................................. 30

Generative Language Models and Open Notes: Exploring the Promise and Limitations *(e51183)*
Charlotte Blease, John Torous, Brian McMillan, Maria Hägglund, Kenneth Mandl. .......................... 34

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. .......................................................... 101

**Editorial**

ChatGPT in Medical Education: A Precursor for Automation Bias? *(e50174)*
Tina Nguyen. .......................................................................................................................... 123

**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. .......................................................... 129
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, with minimal chances of recovery.

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
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 (i.e., 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).
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 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>

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><strong>80-year-old</strong></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><strong>62-year-old</strong></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><strong>Undetermined</strong></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><strong>Both unqualified</strong></td>
<td></td>
</tr>
<tr>
<td>Both are unqualified for ECMO$^a$ per guidelines</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>

$^a$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 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.
References


Abbreviations

ECMO: extracorporeal membrane oxygenation
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–Philipps³, 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 Mahling¹,³, 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 pretrained 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.
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>

**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 specific 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]"

    "Ferdinand Wunderlich:"
    }
```

**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>
<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 inform-ation | 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 inform-ation | 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.” |
<table>
<thead>
<tr>
<th>Aspect and categories</th>
<th>Example(s)</th>
</tr>
</thead>
</table>
| -1: Answer rather implausible | GPT: “I do not follow any specific diet or way of eating. In principle, I eat quite balanced.”
| | GPT: “I used to exercise regularly, but in the past few months, I have been unable to do any sports due to my complaints.” |
| Comment: Both statements do not fully align with the predefined role of a patient with long-standing obesity, but they show aspects of social desirability. |
| -2: Answer very implausible | GPT: “On average, I drink about 10-12 L of fluids per week. This includes water, tea, and occasionally other drinks, such as juice spritzers.” |
| Comment: Here, ChatGPT fails to convert the specified daily fluid intake into a weekly intake. Since polydipsia is an important symptom of hyperglycemia, this statement was rated as very implausible. |

Not applicable

Comment: For QAPs with empty answers due to technical reasons.

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%).
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).

Figure 2. Scatter plot including the trend line for the number of words in the student’s question (x axis) and the number of words in the GPT answer (y axis). Representative variables are displayed as histograms at the top and along the right side. GPT: generative pretrained transformer.
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. For onboarding, the chatbot explained its scope and purpose well by 85.7% (n=14), with 7.1% strongly agreeing. Lastly, 12 (43%) respondents found the chatbot responses useful, appropriate, and informative, while 80.3% (n=15) disagreed with the chatbot failing to recognize a lot of their inputs.
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
Patients, Doctors, and Chatbots

Thomas C Erren1, MPH, MD
Institute and Policlinic for Occupational Medicine, Environmental Medicine and Prevention Research, University Hospital of Cologne, University of Cologne, Köln (Zollstock), Germany

Corresponding Author:
Thomas C Erren, MPH, MD
Institute and Policlinic for Occupational Medicine, Environmental Medicine and Prevention Research
University Hospital of Cologne
University of Cologne
Berliner-Kölnische Allee 4
Köln (Zollstock), 50937
Germany
Phone: 49 022147876780
Fax: 49 022147876795
Email: tim.erren@uni-koeln.de

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.

(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.

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 instance, 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 written 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
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.

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...
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 physicians in writing notes efficiently and clearly, preserving the necessary clinical details. These programs also encourage students and clinicians to write sensitively and empathically, removing loaded jargon or acronyms that may be perceived as offensive (eg, “follow-up” instead of “F/U,” or “shortness of breath” instead of “SOB”) [14,16]. Notably, however, calls for curricular adaptations are isolated, perhaps reflecting wider uncertainty about ORA among the medical community and the perception that the innovation has been foisted on them.

Similarly, interventions to advise patients about how to engage with ORA appear limited [14,36]. This may be owed to a fear among clinicians that encouraging access to web-based records may exacerbate patient anxiety, lead to increased contact time, or risk disagreements and requests to change documentation. We observe that current recommendations in the published and gray literature offer advice on the benefits and risks of accessing ORA, how to maintain password or portal security, and how to discuss errors or disagreements in their notes with clinicians [14,36].

Combined, these clinician and patient support strategies are valuable but have inherent limitations. Training interventions may be variously implemented and take time to become established in mainstream medical education. Even beyond mainstream inclusion of training in medical curricula, it will also be necessary to target the so-called “hidden curriculum”—the set of unspoken and implicit rules and values that trainees may pick up from their mentors and colleagues within clinical practice [37]. It is unclear whether even those strategies that attempt to convert senior or experienced doctors 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 health literacy needs. We suspect that the trade-off may lead to clinicians writing notes that are more suited to a readership like them—individuals with higher health literacy and more years of formal education.

Similarly, 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 disintegrate 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


https://mededu.jmir.org/2024/1/e51183 JMIR Med Educ 2024 | vol. 10 | e51183 | p.38 (page number not for citation purposes)


https://mededu.jmir.org/2024/1/e51183


36. Hannan A. Do you want to see what your doctor or nurse has written about you or check your GP Electronic Health Record? Haughton Thornley Medical Centres. URL: http://tinyurl.com/2vy8rvyk [accessed 2023-12-19]


Abbreviations

AI: artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
ORA: online record access
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.
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, we 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 related to 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 ($\rho=-0.069; P=.003$) while not yielding significance for ChatGPT 4 ($\rho=.87$). 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 ($\rho=-0.289; P<.001$). This was reproducible in ChatGPT 4 ($\rho=-0.344; P<.001$). 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 (ρ=−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 3.5, 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
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 survey 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 \( n = \left[ Z^2 \times p \times (1-p) / E^2 \right] \) [23]. 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 ($H_0$) asserting no association between the variables (use of AI-based applications and country of study) and an alternative hypothesis ($H_1$) 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 1. Demographic characteristics of medical students (n=487).

<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?</td>
<td></td>
</tr>
<tr>
<td>a,b</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?</td>
<td></td>
</tr>
<tr>
<td>a,b,c</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>

\(^a\)Question 5 was exclusively displayed to participants who responded to question 4 with “yes.”

\(^b\)An explanation of the contents of Q5 and Q7 is provided in the text.

\(^c\)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></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></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></td>
</tr>
<tr>
<td>No</td>
<td>69 (23.1)</td>
</tr>
<tr>
<td>Q11: Approximately how many hours have you used the AI-based (chat) application in the last week (7 d)c; (n=189)</td>
<td></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></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></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>S1: ...positively change medicine</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2: ...find useful applications in medicine</td>
<td></td>
<td></td>
<td>.002</td>
<td>3.101</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3: ...influence the choice of my medical specialization</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4: ...reduce the number of jobs for medical staff</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S5: ...improve the quality of patient care</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6: ...improve the process of diagnosis</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S7: ...improve the process of therapy selection</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8: ...negatively affect the doctor-patient relationship</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S9: ...lead to a dehumanization of medicine</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S10: ...negatively affect patient autonomy</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S11: ...negatively affect the autonomy of medical staff</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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S12: ...bring new ethical challenges</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></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 \((P=0.002)\) 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><strong>S13:</strong> ...should be part of medical education</td>
<td></td>
<td></td>
<td></td>
<td></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></td>
<td></td>
</tr>
<tr>
<td><strong>S14:</strong> ...in medical education is adequate</td>
<td></td>
<td></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></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><strong>S15:</strong> ...should include practical content (e.g., exercises to apply AI) in addition to theoretical aspects</td>
<td></td>
<td></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></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><strong>S16:</strong> ...should be based on case studies and application scenarios of AI in medicine</td>
<td></td>
<td></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></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>S17:</strong> ...is an important prerequisite for medical practice</td>
<td></td>
<td></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></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>S18:</strong> ...should be available for medical staff even after graduation</td>
<td></td>
<td></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></td>
<td></td>
</tr>
<tr>
<td>Subgroup 2: no previous use of AI</td>
<td>4 (3-5)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S19:</strong> ...should be updated regularly to reflect advances in AI technology</td>
<td></td>
<td></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></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>S20:</strong> ...is of interest to me</td>
<td></td>
<td></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></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 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><strong>S13: Should be part of medical education</strong></td>
<td>5 (4-5)</td>
<td>5</td>
<td>.37</td>
<td>−0.903</td>
</tr>
<tr>
<td></td>
<td>4 (4-5)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S14: In medical education is adequate</strong></td>
<td>2 (2-3)</td>
<td>2</td>
<td>.21</td>
<td>−1.263</td>
</tr>
<tr>
<td></td>
<td>2 (1-2)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S15: Should include practical content (e.g., exercises to apply AI) in addition to theoretical aspects</strong></td>
<td>4 (0)</td>
<td>4</td>
<td>.80</td>
<td>−0.254</td>
</tr>
<tr>
<td></td>
<td>4 (0)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S16: Should be based on case studies and application scenarios of AI in medicine</strong></td>
<td>4 (3-4)</td>
<td>4</td>
<td>.48</td>
<td>−0.707</td>
</tr>
<tr>
<td></td>
<td>4 (2.5-4.5)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S17: Is an important prerequisite for medical practice</strong></td>
<td>4 (3-4)</td>
<td>4</td>
<td>.90</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>4 (2-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S18: Should be available for medical staff even after graduation</strong></td>
<td>4 (3-4)</td>
<td>4</td>
<td>.17</td>
<td>−1.359</td>
</tr>
<tr>
<td></td>
<td>4 (2-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S19: Should be updated regularly to reflect advances in AI technology</strong></td>
<td>4 (3-4)</td>
<td>4</td>
<td>.17</td>
<td>−1.381</td>
</tr>
<tr>
<td></td>
<td>4 (3-4)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S20: Is of interest to me</strong></td>
<td>4 (3-4)</td>
<td>4</td>
<td>.05</td>
<td>1.782</td>
</tr>
<tr>
<td></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>
<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></td>
<td></td>
</tr>
<tr>
<td>Subgroup 1: previous use of AI</td>
<td>4 (4-5)</td>
<td>5</td>
<td>.22</td>
<td>−1.215</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 the 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 study [28]. It is noteworthy that academic literature on 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 of AI 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
educational frameworks. Cultural attitudes toward AI
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
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
demands 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.

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.

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.

[PDF File (Adobe PDF File), 304 KB - mededu_v10i1e51247_app1.pdf ]

References
   [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
   [FREE Full text] [doi: 10.5858/arpa.2018-0147-OA] [Medline: 30295070]


27. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-10-25]


Abbreviations

AI: artificial intelligence
CPS: clinical practical stage
CS: clinical stage
PCS: preclinical stage
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.
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.

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 κ 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 κ 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...
comprehensive. 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>Partial(^a), 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>

\(^a\)Partial 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/reported) 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><strong>Preparticipation screening (8/24)</strong></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><strong>Frequency (2/11)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fibromyalgia</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>
</tbody>
</table>
ACSM recommends a frequency of 3 days per week [1].

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].

ACSM recommends gradual progression of 4-5 to 8-12 repetitions and increasing from 1 to 2-4 sets per muscle group [1].

ACSM recommends initial intensity should be moderate, progressing to vigorous for greater health benefits [1].

ACSM recommends initial intensity should be moderate, progressing to vigorous for greater health benefits [1].

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].

ACSM recommends initial intensity should be moderate, progressing to vigorous for greater health benefits [1].
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>

aCOPD: 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]. Sarraj 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 health 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 triggers 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...
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 exercise prescription to lower blood pressure”), engaging in multiple or chain prompting and specifically prompting for exercise prescription to lower blood pressure”), engaging in multiple or chain prompting and specifically prompting for exercise prescription to lower blood pressure”). 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 context 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 tools 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).

Multimedia Appendix 2
Summary of major themes derived from artificial intelligence–generated exercise recommendations.

References


34. 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]


40. Zaleski et alJMIR MEDICAL EDUCATION

https://mededu.jmir.org/2024/1/e51308

JMIR Med Educ 2024 | vol. 10 | e51308 | p.83

(page number not for citation purposes)


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/fit.0000000000000843]


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

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 JMIIR 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.
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 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 Long1, MD, MASc; Kayle Lowe2, BSc; Jessica Zhang2, BSc; André dos Santos3, PhD; Alaa Alanazi1, MD; Daniel O’Brien1, MD; Erin D Wright1, MDCM, MEd; David Cote1, 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 otorhinolaryngology/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 otorhinolaryngology/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 otorhinolaryngology 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 AI⁵ 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 OHNS⁶ 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. |

⁵AI: artificial intelligence.  

**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 ($F=0.06$, $F_{\text{crit}}=3.15$; $P=.93$).

**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. 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](https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1) [accessed 2023-05-24]


16. Format of the examination in vascular surgery. Royal College of Physicians and Surgeons of Canada. URL: [http://tinyurl.com/5n8b7jfj](http://tinyurl.com/5n8b7jfj) [accessed 2023-04-08]


18. GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses. OpenAI. URL: [https://openai.com/product/gpt-4](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

©Cai Long, Kayle Lowe, Jessica Zhang, André dos Santos, Alaa Alanazi, Daniel O'Brien, Erin D Wright, David Cote. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.01.2024. 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.
Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project

Nicholas I-Hsien Kuo\(^*\), PhD; Oscar Perez-Concha\(^*\), PhD; Mark Hanly\(^1\), PhD; Emmanuel Mnatzaganian\(^2\), MSc; Brandon Hao\(^2\), BA; Marcus Di Sipio\(^3\), BHSc; Guolin Yu\(^2\), BA; Jash Vanjara\(^5\), MD; Ivy Cerelia Valerie\(^6\), MD; Juliana de Oliveira Costa\(^3\), PhD; Timothy Churches\(^4,5\), MBBS; Sanja Lujic\(^1\), PhD; Jo Hegarty\(^6\), BIT; Louisa Jorm\(^1\), PhD; Sebastiano Barbieri\(^1\), PhD

\(^1\)Centre for Big Data Research in Health, The University of New South Wales, Sydney, Australia
\(^2\)The University of New South Wales, Sydney, Australia
\(^3\)Medicines Intelligence Research Program, School of Population Health, The University of New South Wales, Sydney, Australia
\(^4\)School of Clinical Medicine, University of New South Wales, Sydney, Australia
\(^5\)Ingham Institute of Applied Medical Research, Liverpool, Sydney, Australia
\(^6\)Sydney 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.
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/mm$^3$ 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/mm$^3$ 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 1. The variables of antiretroviral therapy in the HIV data set.

<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&lt;sup&gt;b&lt;/sup&gt; + TDF&lt;sup&gt;c&lt;/sup&gt;, 3TC&lt;sup&gt;d&lt;/sup&gt;, ABC&lt;sup&gt;e&lt;/sup&gt;, FTC + TAF&lt;sup&gt;f&lt;/sup&gt;, DRV&lt;sup&gt;g&lt;/sup&gt; + FTC + TDF, FTC + RTVB&lt;sup&gt;h&lt;/sup&gt; + TDF, other</td>
</tr>
<tr>
<td>Complementary integrase inhibitor (Comp INI)</td>
<td>categorical</td>
<td>N/A</td>
<td>DTG&lt;sup&gt;i&lt;/sup&gt;, RAL&lt;sup&gt;j&lt;/sup&gt;, EVG&lt;sup&gt;k&lt;/sup&gt;, not applied</td>
</tr>
<tr>
<td>Complementary nonnucleoside reverse transcriptase inhibitor (Comp NNRTI)</td>
<td>categorical</td>
<td>N/A</td>
<td>NVP&lt;sup&gt;l&lt;/sup&gt;, EFV&lt;sup&gt;m&lt;/sup&gt;, RPV&lt;sup&gt;n&lt;/sup&gt;, not applied</td>
</tr>
<tr>
<td>Extra protease inhibitor (Extra PI)</td>
<td>categorical</td>
<td>N/A</td>
<td>DRV, RTVB, LPV&lt;sup&gt;o&lt;/sup&gt;, RTV&lt;sup&gt;p&lt;/sup&gt;, ATV&lt;sup&gt;q&lt;/sup&gt;, 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)&lt;sup&gt;r&lt;/sup&gt;</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>

<sup>a</sup>N/A: not applicable.
<sup>b</sup>FTC: emtricitabine.
<sup>c</sup>TDF: tenofovir disoproxil fumarate.
<sup>d</sup>3TC: lamivudine.
<sup>e</sup>ABC: abacavir.
<sup>f</sup>ABC: abacavir.
<sup>g</sup>TAF: tenofovir alafenamide.
<sup>h</sup>DRV: darunavir.
<sup>i</sup>DTG: dolutegravir.
<sup>j</sup>RAL: raltegravir.
<sup>k</sup>EVG: elvitegravir.
<sup>l</sup>NVP: nevirapine.
<sup>m</sup>EFV: efavirenz.
<sup>n</sup>RPV: rilpivirine.
<sup>o</sup>LPV: lopinavir.
<sup>p</sup>RTV: ritonavir.
<sup>q</sup>ATV: atazanavir.
<sup>r</sup>(M): measured.
Figure 2. Inspecting the antiretroviral therapy for an HIV data set (output of the code in Figure 3).

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/mm³ [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.
Figure 5. Time-to-event estimation of viral load suppression for viral load lower than 1000 copies/mL. Red indicates those initiating treatment with FTC + TDF (emtricitabine + tenofovir disoproxil fumarate) and blue for those initiating treatment with 3TC + ABC (lamivudine + abacavir).

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.

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=Dominest](https://www.unsw.edu.au/study/postgraduate/graduate-certificate-in-health-data-science?studentType=Dominest) [accessed 2023-12-27]


35. Consolided guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach, 2nd ed. World Health Organization. URL: [https://www.who.int/publications/i/item/9789241549684](https://www.who.int/publications/i/item/9789241549684) [accessed 2023-12-27]


43. Kuo NIH. The Heath 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](https://figshare.com/articles/dataset/The_Health_Gym_v2_0_Synthetic_Antiretroviral_Therapy_ART_for_HIV_Dataset/22827878) [accessed 2023-12-27]


46. 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](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?cg_plac=&studentType=Domestic [accessed 2023-12-27]

55. R: a language and environment for statistical computing. R-Project. URL: https://www.r-project.org [accessed 2023-12-27]


86. Datasets - CHeReL. Centre for Health Record Linkage. URL: https://www.cherel.org.au/datasets [accessed 2023-12-27]
89. Introducing ChatGPT. OpenAI. URL: 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
Original Paper

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

Introduction

Recently, advances in artificial intelligence (AI) models, more specifically natural language processing (NLP), led to the development of large language models (LLMs) that have shown remarkable performance on a variety of tasks [1-3]. ChatGPT is among the most popular of these models. It was developed by OpenAI and has had several version updates since its inception. GPT-3.5 was among the earlier versions developed, followed by GPT-4.0, developed on March 15, 2023, as a more robust, concise, and intelligent model. ChatGPT has become
quite famous for its outstanding ability to answer questions and assist in many tasks [4].

Medical education relies heavily 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 both GPT-4.0 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).

<table>
<thead>
<tr>
<th>Question: What medical discipline deals with conditions of the eye</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Dermatology</td>
</tr>
<tr>
<td>B. Endocrinology</td>
</tr>
<tr>
<td>C. Ophthalmology</td>
</tr>
<tr>
<td>D. Rheumatology</td>
</tr>
</tbody>
</table>

**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, 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>ChatGPT’s performance (correct answers), n (%)</th>
<th>Human performance (correct answers), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Board of Ophthalmology difficulty level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>49 (76)</td>
<td>19 (100)</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>73 (59)</td>
<td>43 (91)</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>35 (56)</td>
<td>38 (84)</td>
<td>53</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>10 (59)</td>
<td>37</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>4 (66.67)</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-4.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Board of Ophthalmology difficulty level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>19 (100)</td>
<td>34 (53)</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>2 (44.26)</td>
<td>1 (88)</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>3 (28)</td>
<td>28 (51.6)</td>
<td>53</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>12 (60)</td>
<td>37</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>3 (50)</td>
<td>26</td>
</tr>
</tbody>
</table>

- \( P = .04 \) on comparing the performance of GPT-4.0 across different difficulty levels.
- \( P = .003 \) on comparing the performance of GPT-4.0 across different difficulty levels.
- \( P = .49 \) on comparing the performance of GPT-3.5 across different difficulty levels.
- \( P = .18 \) on comparing the performance of GPT-3.5 across different difficulty levels.
- 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, n (%)</th>
<th>Topic</th>
<th>Correct answers by GPT-3.5, 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>

- \( P = .02 \) for differences in the number of correct answers provided by GPT-4.0 among different categories.
- \( P = .03 \) for differences in the number of correct answers provided by GPT-3.5 among different topics.
- \( P < .05 \).

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 to answer correctly than topics such as oncology and pathology, which rely more on clinical knowledge. Taloni et al [8] reported a better performance of ChatGPT on clinical rather than surgical cases.

The moderate accuracy of ChatGPT-3.5 has been widely replicated in various studies. Gilson et al [7] found accuracies ranging between 42% and 64.4% in USMLE steps 1 and 2 examinations, numbers similar to those noted in this study [7]. The paper also records a decrease in the proportion of correct answers as difficulty level increases, which has been noted in this study as well. Another study by Huh [9] showed that ChatGPT’s performance was significantly lower than that of Korean medical students in a parasitology examination. A letter to the editor of the journal Resuscitation revealed that ChatGPT did not reach the passing threshold for the Life Support examination [10]. The cited studies indicate the moderate capabilities of ChatGPT in answering clinically related questions. More studies are needed to show how we can best optimize ChatGPT for medical education. Mihalache et al [11] assessed the performance of ChatGPT on the OKAP and found that it provided 46% correct answers, not unlike the proportion of OB-WQE questions correctly answered by GPT-3.5 in this study. All the aforementioned studies used ChatGPT-3.5 in their analysis. More recent studies have assessed the efficacy of ChatGPT-4.0. A study by Lim et al [12] assessed the performance of GPT-4.0 on myopia-related questions, and the model performed with 80.6% adequate responses, compared to 61.3% for GPT-3.5. Taloni et al [8] assessed the use of ChatGPT-4.0 and ChatGPT-3.5 in the American Academy of Ophthalmology’s self-assessment questions; their study found that GPT-4.0 (82.4% of correct answers) performed better than both humans (75.7% of correct answers) and GPT-3.5 (65.9% of correct answers). The study also assessed the performance of these models across various topics [8]. Similar to our results, Taloni et al [8] found that ChatGPT performed better on ocular oncology and pathology 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]
Editorial

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: nguy.t921@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, we must consider 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 overly rely 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 overrelied 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 care, 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 Ridder 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 Monahan¹, MD; Edward Gould¹, MD; Todd Rice¹, MD; Patty Wright¹, MD; Eduard Vasilevskis¹, MD; Frank Harrell¹, PhD; Monique Drago¹, EdD; Sarah Mitchell¹, 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=.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<.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</th>
<th>Value at the end of the observation (range during observation period)</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 (prepandemic)</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(b), mean (SD)</strong></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>

\(a\) N/A: not applicable.

\(b\) 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\%).
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 professor (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.
**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

Edited by T Leung, T de Azevedo Cardoso; submitted 20.10.22; peer-reviewed by S Hertling, M Hedges; comments to author 20.03.23; revised version received 23.03.23; accepted 19.07.23; published 03.01.24.

Please cite as:
Impact of the COVID-19 Pandemic on Medical Grand Rounds Attendance: Comparison of In-Person and Remote Conferences
JMIR Med Educ 2024;10:e43705
URL: https://mededu.jmir.org/2024/1/e43705
doi: 10.2196/43705
PMID: 38029287

https://mededu.jmir.org/2024/1/e43705